2012

CHEMOMETRIC ANALYSIS OF COMPREHENSIVE TWO-DIMENSIONAL LIQUID CHROMATOGRAPHIC-DIODE ARRAY DETECTION DATA: PEAK RESOLUTION, QUANTIFICATION AND RAPID SCREENING

Hope P. Bailey
Virginia Commonwealth University

CHEMOMETRIC ANALYSIS OF COMPREHENSIVE TWO-DIMENSIONAL LIQUID CHROMATOGRAPHIC-DIODE ARRAY DETECTION DATA: PEAK RESOLUTION, QUANTIFICATION AND RAPID SCREENING

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

by

Hope Patricia Bailey
B.S., Virginia Commonwealth University, 2004

Director: Sarah C. Rutan
Professor, Department of Chemistry

Virginia Commonwealth University
Richmond, Virginia
December2012

Acknowledgement

This has been a journey I had not expected to undertake and has been one of the most fulfilling accomplishments of my life. At the beginning of this road, I believed God had a plan even though I had no way of seeing just how far He planned for me to go and where I was to end up. During this adventure, I have had two wonder boys, Joshua and Caleb, they are the BEST accomplishment my life will ever have and I pray that one day the example to try even if you think you cannot and to never give up will help them to accomplish more than I ever could. Many people have enabled me, over the past several years, to come to the final conclusion. My husband, Paul, and my mom and dad, Daleen and Glenn Purdie, offered assistance with watching my boys when I needed the additional time for school. Many friends provided emotional sanity (Melinda, Tasha, Laurie…), ridiculous amounts of tech support (Zack and Mark) or much needed commiseration (Daniela, Sarah, Railey). All of the undergrads and graduate students that have come and gone in the Rutan lab have made the time spent in the lab productive, fun and a great learning environment (most importantly the "sociopath" of course…LOL, Robert). But I would never have even started on this adventure if not for Dr. Rutan; and I certainly would not have learned or accomplished this much without her patience with me, quiet confidence in me and the occasional much needed and deserved swift kick. So very thankful I made this journey with so many great friends, family and the best adviser ever!

Table of Contents

Acknowledgements.....ii

List of Figures.....vii

List of Tables.....xiv

List of Abbreviations.....xvi

List of Symbols.....xix

Abstract.....xxiii

Chapter 1: Overview of Objectives.....1

Chapter 2: Comprehensive Two-Dimensional Liquid Chromatography (LC × LC).....7

2.1 Instrumentation.....7

2.2 Two-Dimensional Liquid Chromatography.....9

2.3 High Temperature in Fast LC × LC.....11

2.4 Peak Capacity.....15

2.5 Possible Detectors.....20

2.6 Peak Quantification.....21

Chapter 3: Chemometric Techniques and Theory.....26

3.1 Data Structure.....27

3.2 Singular Value Decomposition (SVD).....30

3.3 Iterative Key Set Factor Analysis (IKSFA).....32

3.4 Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS).....34

3.5 Parallel Factor Analysis (PARFAC).....40

Chapter 4: Applicability to and Chromatographic Separation of Metabolomics Samples.....43

4.1 Urine and Standards Mixture.....44

4.2 Wine.....47

4.3 Phenytoin.....50

Chapter 5: Chemometric Resolution and Quantification of Four-Way Data Arising from Comprehensive 2D-LC-DAD Analysis of Human Urine.....54

5.1 Quantification Algorithm Development (relative concentration determination).....54

5.2 Data Analysis Scheme.....57

5.3 Standards Mixture Analysis.....60

5.4 Urine Control Sample Analysis.....63

5.5 Comparison to Previous Rutan Group Work.....67

5.6 Data Analysis Considerations.....68

5.7 Conclusions.....72

Chapter 6: Chemometric Analysis of Targeted 3D-LC-DAD Data for Accurate and Precise Quantification of Phenytoin in Wastewater Samples.....77

6.1 IKSFA-ALS-ssel.....77

6.2 IKSFA-ALS with All Constraints 80

6.3 Statistical Analysis.....82

6.4 sLC × LC Importance.....84

6.5 Conclusions.....85

Chapter 7: Factors That Affect Quantification of Diode Array Data in Comprehensive Two-Dimensional Liquid Chromatography using Chemometric Data Analysis 87

7.1 Review of the Implemented Chemometric Method 88

7.2 Comparison of Quantification Methods 89

7.3 Overview of the Data Analysis Method 91

7.4 Spectral and Chromatographic Rank Deficiencies 94

7.5 Retention Time Shifts 97

7.6 Dynamic Range Issues 100

7.7 Inadequate Background Removal 103

7.8 Additional issues 104

7.9 Conclusions 107

Chapter 8: Comparison of Chemometric Methods and Statistical Analysis for the Screening of Comprehensive Two-Dimensional Liquid Chromatographic Analysis of Wine.....110

8.1 Theory

8.1.1 Alignment.....111

8.1.2 Similarity Index (SI).....112

8.1.3 Fisher Ratio (FR).....114

8.2 Experimental

8.2.1 Data organization.....115

8.2.2 Data analysis scheme.....117

8.3 Results and Discussion.....119

8.3.1 Statistical analysis of concentration data.....120

8.3.2 Threshold determinations.....123

8.3.3 Geographical variability (Similarity Index method).....128

8.3.4 Geographical variability (Fisher Ratio method).....130

8.3.5 Simulated data analysis.....131

8.4 Conclusions.....134

Chapter 9: Conclusions and Future Work.....136

9.1 Goal of Resolution.....137

9.2 Goal of Quantification.....138

9.3 Goal of Rapid Screening.....139

9.4 Future Work.....140

References.....143

Appendix.....153

Vita.....164

List of Figures

Chapter 2:

Figure 2.1: Schematic diagram of a liquid chromatograph. MP1, MP2 and MP3 are three different solutions that can be combined in the mixing vessel in differing ratios to be passed through the chromatographic instrumentation via a high pressure pump. The sample to be analyzed is then injected into the system such that the mixture compounds are separated while traveling through the analytical column. The results are displayed as a chromatogram that plots the detector response as a function of elution time.……..……...8

Figure 2.2: Chromatogram of urine sample resulting from a 30 minute 1D-LC separation. The chromatographic conditions are as follows: gradient elution from 0 to 70% B from 0 to 23 min, where A is 20 mM sodium phosphate, 0.1 mM EDTA at pH 6 and B is acetonitrile, with a flow rate of 0.1 mL/min. The stationary phase is a lab-made hydroxylated-hypercrosslinked material [15] packed in a 200 mm x 1.0 mm column……...9

Figure 2.3: Contour plot of an LC × LC separation of the same urine sample shown in Figure 2.2 and the chromatographic conditions for the ^{1}D column are the same as provided in the caption of Figure 2.2. The chromatographic conditions for the ^{2}D column are as follows: gradient elution from 0 to 100% from 0 to 17.45 s, where A is 20 mM phosphoric acid and B is acetonitrile, with a flow rate of 3 mL/min. The re-equilibration time was 3 seconds. Two pumps are used to deliver the samples loaded in the two 35 µL loops in an alternating fashion to the ^{2}D column, where the ^{2}D stationary phase is a carbon-clad zirconia material packed in a 3.3 mm x 2.1 mm column [22]……...10

Figure 2.4: Schematic of LC × LC dual gradient, high temperature ^{2}D system [3] showing the use of three binary pumps, two eluent preheaters, heating jackets around both the ^{1}D and ^{2}D dimension columns, a six port switching valve for the ^{2}D gradient mobile phase and a ten port switching value that collects and delivers the ^{1}D aliquot to the ^{2}D system……..12

Figure 2.5: van Deemter plot for a packed LC column showing the relationship between the plate height (H) of the column and the flow rate of the chromatographic system where the black line represents the total plate height details and the brown, yellow and green lines are the contributions due to the three different rate terms…….13

Figure 2.6: Illustration of two van Deemter curves at different temperatures. The black curve represents the use of a low temperature, while the red curve is represents the system at a higher temperature. Note that the value of H_{min} is unchanged for both curves, but that it now corresponds to a higher a linear velocity for the high temperature curve……15

Figure 2.7: Illustration of peak coverage of the available separation space for a 2D separation system under the conditions of (A) non-orthogonality (B) partial orthogonality and (C) complete orthogonality. The red lines illustrate the concept of $f_{coverage}$ if the minimum convex hull method was applied to the illustrated data…….18

Figure 2.8: (A) Contour plot of a LC × LC peak after background subtraction using LCImage software. The (red) line is the peak boundary as determined by the LCImage software via the watershed algorithm. (B) The corresponding sequence of 2nd dimension chromatograms. (C) The integrated area of slice 5. (D) The integrated area of slice 10. (This peak corresponds to the standards mixture peak 1 as discussed in Chapter 4.)……...24

Chapter 3:

Figure 3.1: Illustration describing the relationship between components and compounds……..28

Figure 3.2: Mesh plot of a 1D chromatographic peak and corresponding spectrum and two individual plots of the corresponding **R** matrix (chromatogram) and **S** matrix (spectrum)……..28

Figure 3.3: Several different types of plots for the 2nd sample of the standards mixture raw data (see Chapter 4). (A) Plot of the two-way raw data as collected by the instrument, such that all 2nd dimension injections of a corresponding sample injection are sequenced end to end. The insert shows an enlarged section of the sequenced data that was determined to encompass the corresponding peak illustrated in the contour plot at 216 nm. (B) Using the determined section dimensions, the same peak is shown as a contour plot at 216 nm in the box (C) along with its corresponding sequenced 2nd dimension chromatograms……...29

Figure 3.4: Visual representation of 4-way data set where the 1st and 2nd dimensions are 1st and 2nd chromatographic retention times respectively, the 3rd dimension is spectral wavelengths and the 4th dimension is the number of injected samples…….....30

Figure 3.5: Visual representation of unfolding four-way data in which the 4th dimension (*i.e.,* the spectral dimension) is conserved during the unfolding of the four-way data to two-way data…….30

Figure 3.6: Scree plot of the relative importance of the singular values plotted as a function of the number of factors in the data matrix…….32

Figure 3.7: Visual representation of MCR-ALS data decomposition of 3-way data to 2-way data [11]……..36

Figure 3.8: Constraints associated with MCR-ALS. From top to bottom: non-negativity, unimodality, trilinearity and spectral selectivity constraints. From left to right: before and after application of the corresponding constraint……...37

Figure 3.9: Schematic illustration of the resolution results from the IKSFA-ALS-ssel analysis in which both the data structure and corresponding graphical representations of the resolved spectra and chromatograms are shown. The data structure as shown consists of *J*=7 2nd dimension slices and K samples (1st dimension injections). Panel 1 illustrates the resolved spectral matrix. The dimensions of the resolved S^T data matrix in this example are four spectral components by the total number of wavelengths collected (*L*). The corresponding spectra for each of the four components are plotted such that the y-axis is relative intensity versus wavelength. Panel 2 illustrates the resolved chromatographic matrix. Recall that each 1st chromatographic data point is equal to a 2nd dimension slice

and that a 2nd dimension slice consists of I data points. The dimensions of the resolved C data matrix is unfolded to combine the 1st chromatographic dimension (J), the 2nd chromatographic dimension (*I*) and the number of samples (*K*) by the four spectral components. The four spectral components are color coordinated to correspond to their chromatographic counterpart. Each resolved chromatographic peak is graphically represented in two ways (1) by a contour plot of the 1st chromatographic dimension by the 2nd chromatographic dimension plotted for a given wavelength and a given sample and (2) a sequence of 2nd dimension chromatograms. Panel 3 shows the corresponding sequence of 2nd dimension chromatograms of component 1 for all samples (1−*K*)……..39

Chapter 4:

Figure 4.1: Contour plots at 216 nm of two different sample types within the 64 injection 2D-LC-DAD run. (A) Contour plot of the third replicate injection of the standards mixture where peak 1 is indole-3-acetonitrile, peak 2 is indole-3-propionic acid, peak 3 is indole-3-acetic acid, peak 4 is tryptophan peak 5 is hydroxytryptophan and peak 6 is tyrosine. (B) Contour plot of the seventh replicate injection of the urine control standard. Inset shows the section of data selected for chemometric analysis……...47

Figure 4.2: Contour plot of HRC C 1st replicate at 216 nm. The boxed area is the section of the chromatograms analyzed in this work…….50

Figure 4.3: Chromatograms observed at the outlet of each dimension of separation in the 3D-LC/UV system for the separation of the 1000-fold concentrated WWTP effluent sample. The red chromatogram: neat WWTPE extract, blue chromatogram: phenytoin and chlorophene standards spiked into the WWTPE at 500 and 50 ppb, black chromatogram: phenytoin and chlorophene standards spiked in DI water at 500 and 50 ppb. (reproduced from reference [101] with permission from Elsevier…….……...52

Figure 4.4: Contour plots of various sample injections at 216 nm for the phenytoin study before chemometric analysis. The shaded portion of the plots is the section of the data eliminated from the chemometric analysis of the data. (A) Contour plot of DI water sample spiked with 25 ppb phenytoin. (B) Contour plot of the WWTPE sample without a spiked amount of phenytoin. (C) Contour plot of the WWTPE sample spiked with 150 ppb phenytoin……53

Chapter 5:

Figure 5.1: Comparison of the subsection for the standards mixture containing Peak 1 showing the sequence of resolved 2nd dimension chromatograms and the raw data for the injection of six replicate samples onto the 1st dimension column. (A) The data after application of the developed chemometric method. The line drawn under each 2nd dimension peak shows the manually determined baseline, and the areas for each of the second dimension peaks (shown at the top of the peaks) are totaled (shown at the bottom of each peak grouping), giving the relative concentrations of Peak 1 for each of the six sample injections and the % RSD showing the precision of the quantification. (B) A plot of the sequenced second dimension chromatograms of the raw data. Each 1st dimension sample injection gives rise to seven 2nd dimension injections

with four of those injections containing the peak of interest and three injections consisting only of the background in this example…….……....56

Figure 5.2: Chemometric data analysis scheme used in the resolution and quantification of LC × LC data……58

Figure 5.3: Contour plots at 216 nm of two different sample types within the 64 injection 2D-LC-DAD run. (A) Contour plot of the third replicate injection of the standards mixture where peak 1 is indole-3-acetonitrile, peak 2 is indole-3-propionic acid, peak 3 is indole-3-acetic acid, peak 4 is tryptophan peak 5 is hydroxytryptophan and peak 6 is tyrosine. (B) Contour plot of the seventh replicate injection of the urine control standard. Inset shows the section of data selected for chemometric analysis……..……60

Figure 5.4: Contour plot of the 7th urine control at 216 nm showing 34 resolved peaks. The N preceding 16 of the 34 resolved and quantified peaks signifies that those peaks were found and resolved only after application of the developed chemometric method (newly found) while the other 18 peaks were visually observable prior to chemometric analysis. The two bar graphs show % RSD values calculated for the corresponding peaks. The star on the visually observed peaks graph indicates that Peak 8 is considered to be a chemically unstable compound……..…..64

Figure 5.5 (A) Contour plot of a subsection of urine control data at 216 nm in which resolution and quantification of Peak N16 is the goal for the chemometric analysis. (B) The chromatographic and corresponding spectral results for each component of the 8 component IKSFA-ALS-ssel analysis for the above subsection of raw data……...….65

Figure 5.6: Overlaid contour plot of maize data analyzed by Porter *et al.* [2]. The blue contour plot is the first injection of the mutant sample, the green contour plot is the indole standard mixture and the red contour plot in the second injection of the wild-type sample. The inset corresponds to the outlined section …….…..…67

Figure 5.7: Component spectra (black) and background spectra (red) before (A) and after (B) implementation of the spectral selectivity and spectral non-negativity constraints. By zeroing the chemical component spectra after 440 nm, the algorithm is better able to resolve the background spectra from the compound spectra…….....71

Chapter 6:

Figure 6.1: Contour plots of various sample injections at 216 nm before chemometric analysis. The shaded portion of the plots is the section of the data eliminated from the chemometric analysis of the data. (A) Contour plot of DI water sample spiked with 25 ppb phenytoin. (B) Contour plot of the WWTPE sample without a spiked amount of phenytoin. (C) Contour plot of the WWTPE sample spiked with 150 ppb phenytoin……77

Figure 6.2: Chromatographic results of the chemometric analysis for a six component model for the 75 ppb phenytoin standard sample and the 75 ppb phenytoin in addition to the WWTPE sample. (A) Analysis without implementation of the sample selectivity constraint which overfits the DI water samples and assigns some of the phenytoin peak to component 5 (as indicated by the arrow), the interferent component in the WWTPE samples. (B) Analysis with implementation of the sample selectivity constraint such that the concentrations of components 1 and 5 of the DI water samples were constrained to be zero……78

Figure 6.3: Plots showing the overlap of the phenytoin and interferent peaks. (A) Contour plot of the fifth IKSFA-ALS-ssel component for the 150 ppb spiked WWTPE sample which shows the incorrect assignment of a portion of the phenytoin peak eluting after the interferent peak in the third retention time dimension. (B) Overlay of the 150 ppb spiked WWTPE sample for three third dimension sequenced chromatograms, such that the blue (bottom) series of chromatograms is the raw data, the green (middle) series of chromatograms is the IKSFA-ALS-ssel analyzed data for the fifth component, which shows the incomplete resolution of the phenytoin and interferent peaks and the red (top) series of chromatograms is the IKSFA-ALS-ssel-csel result for the fifth component, which shows complete resolution of the phenytoin from the interferent peak……....79

Chapter 7:

Figure 7.1: Contour plot of the urine control at 216 nm showing 34 resolved peaks. The N preceding 16 of the 34 resolved and quantified peaks signifies that those peaks were found and resolved only after application of the developed chemometric method (newly found) while the other 18 peaks were visually observable prior to chemometric analysis…….. 92

Figure 7.2: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of peak 13 subsection of urine control data. (B) Chromatographic and spectral IKSFA-ALS-ssel results for a 5 component model. This figure illustrates data that are not rank deficient in either the chromatographic or spectral dimensions. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm. The star denotes a cut off peak that was not analyzed and therefore not assigned a peak number……..94

Figure 7.3: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of urine control data encompassing peaks 6 and 7 which have different 2nd dimension retention times, the same 1st dimension retention time and very similar spectra. The 2 boxes show the 2 different subsections used to separately analyze each of the peaks. (B) Overlay of the corresponding raw spectra for peak 6 (dashed line or red spectrum) and peak 7 (dotted line or blue spectrum) measured at the corresponding peak maxima, illustrating the spectral similarity of peaks 6 and 7. (C) Chromatographic and spectral IKSFA-ALS-ssel results for the component that contained peak 7. (D) Chromatographic and spectral IKSFA-ALS-ssel results for the component that contained peak 6. The chromatographic axis labels in C and D are the same as those in A, and the wavelength range is 200–700 nm……..95

Figure 7.4: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of the subsection for peak 9. (B) Chromatographic and spectral IKSFA-ALS-ssel results showing the unique resolved spectra for peaks 9, N14 and N15 that appear to have the same first and second dimension retention times. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm……..97

Figure 7.5: Illustration of phase shifting of ^{1}D peak. (A) The red peak simulates an exactly in phase first dimension peak having a max centered within slice 3. The yellow and green curves are peaks that have shifted earlier in the retention time but have not shifted to an exactly out of phase position. (B) Histogram representation of the area under the curve for each of the three represented peaks……..98

Figure 7.6: (A) Contour plots at 216 nm of the 1st and 7th sample injections after multivariate analysis of the peak 12 subsection. (B) Overlay of the sequence of 2nd dimension chromatograms after IKSFA-ALS-ssel analysis such that the blue or dashed line chromatogram corresponds to sample injection 1 and the black or solid line chromatogram corresponds to sample injection 7 which shows the coelution of peaks 11 and 12 owing to phase shifting in sample injection 7. (C) Schematic representation of the effects of phase shifting on the quantitative analysis of chromatographically overlapped peaks……..…..100

Figure 7.7: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of peak N16 subsection of the urine control data. (B) Chromatographic and spectral IKSFA-ALS-ssel results for an 8 component model. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm……..……101

Figure 7.8: (A) Contour plots at 216 nm of the IKSFA-ALS-ssel resolved component for the analysis of peak N16 for all 14 sample injections where injection 1 is the top left hand corner and injections follow sequentially to injection 14 in the bottom right hand corner. Injections 6–8 are within the rectangular box. (B) Sequential 2nd dimension chromatograms for sample injection clearly indicating the presence of "embedded" peaks in the 4th and 6th slices as shown by the arrows. There are 61 data points for each 2nd dimension slice and 8 1st dimension data points for a total of 489 data points on the sequenced chromatograms. (C) Corresponding contour plot at 216 nm for injection 1. The chromatographic axis labels in A are the same as those in C, and the wavelength range is 200–700 nm……..…102

Figure 7.9: Peak splitting that results in a negative peak that corresponds to the analyte of interest in the background component. (A) Contour plots of the peak of interest and of the background after multivariate analysis without implementation of the spectral constraints, along with corresponding overlay plots of the sequence of 2nd dimension chromatograms and spectra. (B) Contour plots of the peak of interest and of the background after IKSFA-ALS-ssel, along with corresponding overlay plots of the sequence of 2nd dimension chromatograms and spectra. The dashed curve corresponds to a background component, and the solid curve corresponds to the component of interest……105

Chapter 8:

Figure 8.1: Schematic representation of global alignment applied in the 2nd retention time dimension. (A) Three sample injections illustrated with only retention time shifting in the second dimension. (B) Each sample injection is aligned using a global parameter such that the maximum for each peak is in the same position and the data size dimensionality remains consistent……112

Figure 8.2: Contour plot of HRC C 1st replicate at 216 nm. The boxed area is the section of the chromatograms analyzed in this work……..116

Figure 8.3: (A) Similarity Index contour plot showing each of the nine analyzed peaks and the associated scale from 0 (most dissimilar) to 1 (most similar). (B) Fisher ratio contour plotted using a logarithmic base 10 for scaling showing the nine analyzed peaks and the scale……..117

List of Tables

Table 2.1: Methods used for the quantitative analysis of LC × LC data……...…22

Table 3.1: Quantitative results after resolution of the target analytes in each mixture…….41

Table 5.1: % RSD results for the precision of peak quantification of both raw and IKSFA-ALS-ssel resolved data of the standards mixture injections for Peak 1 through Peak 5…….....61

Table 5.2: Effects of subsection size on the % RSDs of IKSFA-ALS-ssel analyzed standard mixture data for Peak 1–Peak 5……...…63

Table 5.3: Comparison of % RSD values for duplicate samples resulting from PARAFAC-ALS method [1, 2] and IKSFA-ALS methods in the analysis of maize data. NP: the compound was not present in the wild-type samples……...68

Table 6.1: % RSD of the duplicate sample injections for both the DI water and WWTPE samples after chemometric analysis……..80

Table 6.2: Comparison of the unknown sample calculations using both the standard addition and calibration methods for the chemometric method with and without the sample selectivity constraint ssel constraint……..83

Table 7.1: % RSD results for peak quantification of both the raw and IKSFA-ALS-ssel resolved data of the standards mixture injections……..90

Table 7.2: Precision of peak quantification of urine control sample……93

Table 7.3: Combined analysis results of several smaller subsections showing all of the detected peaks that were found in the subsection used for the analysis of peak N16……..106

Table 8.1: Peak statistics, SI and FR values……...122

Table 8.2: Minimum similarity index found for the replicate analysis of the 5 different samples studied……124

Table 8.3: Results of Equivalence Test where E shows equivalence between the 2 samples and NE showed no equivalence……...126

Table 8.4: Results of Tukey's (HSD) test where D shows a statistical difference between the 2 samples and ND shows no difference…….....126

Table 8.5: Peak rankings for the wine data for the SI and FR methods……..128

Table 8.6: Peak rankings for the simulated data consisting of four different background contributions for the SI and FR methods…….133

List of Abbreviations

A

AED	anti-epileptic drug
ATLD	alternating trilinear decomposition

C

CA	cluster analysis
CV	Ceplecha Vineyard

D

DA	discriminate analysis
DAD	diode array detectors
DI	distilled water
DTLD	direct trilinear decomposition

E

ESI	electrospray ionization

G

GC	gas chromatography
GC × GC	comprehensive two dimensional gas chromatography
GS-MS	gas chromatography mass spectrometry

H

HELP	heuristic evolving latent projections
HPLC	high performance liquid chromatography
HRC	horticultural research center

I

IKSFA	iterative key set factor analysis
IKSFA-ALS-ssel	iterative key set factor analysis-alternating least squares with spectral selectivity

L

LC × LC	comprehensive two dimensional liquid chromatography
LC-LC	“heart-cutting” two-dimensional liquid chromatography
LC-MS/MS	liquid chromatography with tandem mass spectrometry

M

MCR-ALS	multivariate curve resolution-alternating least squares
MS	mass spectrometers
M-S-F	Murphy, Schure and Foley sampling criterion

N

NMR	nuclear magnetic resonance

P

PARAFAC	parallel factor analysis
PCA	principal component analysis
PPCPs	pharmaceuticals and personal care products

S

sLC × LC	selective comprehensive two-dimensional HPLC
SOT	statistical overlap theory
SPE	solid phase extraction
SVD	singular value decomposition

U

UHPLC	ultra high performance liquid chromatography
UHPLC-TOF-MS	ultra high performance liquid chromatography-time-of-flight-MS

W

WFA	window factor analysis
WV	Winter Vineyard
WWTPE	wastewater treatment plant effluent

% RSD	percent relative standard deviation
^{1}D	separated first dimension column
^{2}D	second dimension column
2D-HPLC	two dimensional high performance liquid chromatography
3D-LC	three dimensional liquid chromatography

List of Symbols and Variables

A, **B**, **C** and **D**	loadings matrices associated with PARAFAC containing the variables a_{in}, b_{jn}, c_{kn}, d_{ln}
A	eddy diffusion term in the van Deemter equation
B	longitudinal diffusion term in the van Deemter equation
C_S+C_M	resistance to mass transfer term in the van Deemter equation
D_M	diffusion coefficient in the mobile phase
E	error matrix
F	Fisher ratio
H	column plate height
I	number of data points in each 2nd dimension chromatogram
J	number of data points in each 1st dimension chromatogram
L	number of points in each spectrum
K	number of different samples analyzed
M	equal to *IJK* for IKSFA-ALS-ssel analysis
M_2	molecular weight of the solvent
N	number of components comprising the data set **X**
N_p	number of replicates in class *p*

P	equal to IJ for similarity index and Fisher ratio analysis
Q	number of classes used in the Fisher ratio method
$\mathbf{R}$	chromatographic matrix; the columns contain the chromatograms of the individual pure components present in the samples represented by matrix $\mathbf{X}$
R_s	chromatographic resolution
$\mathbf{S}$	spectral matrix; the columns contain the spectra of the individual pure components present in the samples represented by matrix $\mathbf{X}$
SS_{fact}	sum of squares between classes
SS_R	residual sum of squares within classes
T	temperature
$\mathbf{U}$	left singular vectors of $\mathbf{X}$ from SVD analysis
$\mathbf{V}$	right singular vectors of $\mathbf{X}$ from SVD analysis
V	molar volume of the solute
$\mathbf{W}$	singular value matrix from SVD
$\mathbf{X}$	data matrix consisting of absorbance values as a function of elution time (columns-chromatogram) and absorbance values as a function of wavelength (rows-spectra)
$\underline{\underline{\mathbf{X}}}$	data array I×J×K×L
a_{in}	contains the second chromatographic response for component n at the i^{th} data point for the PARAFAC model
b_{jn}	contains the first chromatographic response for component n at the j^{th} first dimension data point for the PARAFAC model
c_{kn}	contains the relative concentration of component n at the k^{th} data point for the PARAFAC model
d_{ln}	contains the spectral response for component n at the i^{th} spectral data point for the PARAFAC model

e_{ijkl}	residual error
d_p	particle diameter
n_c	peak capacity
$n_{c,2D}$	ideal 2D peak capacity
$n'_{c,2D}$	2D peak capacity corrected for under sampling
${}^1n_c$	uncorrected first dimension peak capacity
${}^2n_c$	uncorrected second dimension peak capacity
r_p	correlation coefficient for the p^{th} chromatographic time point (similarity index method)
s_p	similarity index for the p^{th} chromatographic data point
t_R	retention time
$t_{R,\,first}$	retention time of the first peak to elute from the column
$t_{R,\,last}$	retention time of the last peak to elute from the column
t_2 and t_1	retention times of adjacent peaks
u	linear velocity of the mobile phase.
$\tilde{\mathbf{u}}_{key1}$	normalized first key row from SVD matrix **U**
$\tilde{\mathbf{u}}_{key2}$	normalized first key row from SVD matrix **U**
$\mathbf{u_r}$	row vector from the matrix **U**
$\tilde{\mathbf{u}}_r$	normalized row vector from the matrix **U**
$\tilde{\mathbf{u}}_1$	normalized first row vector from the matrix **U**

w_1 and w_2	widths at the base of chromatographic peaks t_1 and t_2
$\bar{x}$	overall mean of the data
x_{np}	n^{th} measurement in the p^{th} class of the data
$\bar{x}_p$	mean of the p^{th} class
$<\beta>$	correction for sampling-induced peak broadening
ϕ	sampling phase
T	center of the sample cycle nearest to the peak maximum
τ	second dimension run time (modulation period)
σ_{cl}^2	class-to-class variance
σ_{err}^2	within class variance
η	viscosity
Ψ	association constant for solvent

Abstract

CHEMOMETRIC ANALYSIS OF COMPREHENSIVE TWO-DIMENSIONAL LIQUID CHROMATOGRAPHIC-DIODE ARRAY DETECTION DATA: PEAK RESOLUTION, QUANTIFICATION AND RAPID SCREENING

By Hope Patricia Bailey, Ph.D.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University

Virginia Commonwealth University, 2012

Director: Sarah C. Rutan, Professor, Department of Chemistry

This research project sought to explore, compare and develop chemometric methods with the goal of resolving chromatographically overlapped peaks though the use of spectral information gained from the four-way data sets associated with comprehensive two-dimensional liquid chromatography with diode array detection (LC × LC-DAD). A chemometric method combining iterative key set factor analysis (IKSFA) and multivariate curve resolution-alternating least squares (MCR-ALS) was developed. In the section of urine data analyzed, over 50 peaks were found, with 18 visually observable and 32 additional compounds found only after application of the chemometric method.

Upon successful chemometric resolution of chromatographically overlapped peaks, accurate and precise quantification was then necessary. Of the compared methods for quantification, the manual baseline method was determined to offer the best precisions. Of the 50 found peaks from the urine analysis, 34 were successfully quantified using the manual baseline method with percent relative standard deviations ranging from 0.09 to 16. The accuracy

of quantification was then investigated by the analysis of wastewater treatment plant effluent (WWTPE) samples. The chemometrically determined concentration of the unknown phenytoin sample was found to not exhibit a significant difference from the result obtained by the LC-MS/MS reference method, and the precision of the IKSFA-ALS method was better than that of the precision of the LC-MS/MS analysis. Chromatographic factors (data complexity, large dynamic range, retention time shifting, chromatographic and spectral peak overlap and background removal, were all found to affect the quantification results.

The last part of this work focused on rapid screening methods that were capable of locating peaks between samples that exhibited significant differences in concentration. The aim here was to reduce the amount of data required to be resolved and quantified to only those peaks that were of interest. This would then reduce the time required to analyze large, complex samples by eliminating the need to first quantify all peaks in a given sample for many different samples. Both the similarity index (SI) method and the Fisher ratio (FR) method were found to fulfill this requirement in a rapid means of screening fifteen wine samples.

Chapter 1: Overview and Objectives

Modern chromatography techniques, such as high performance liquid chromatography (HPLC) and gas chromatography (GC), are increasingly capable of analyzing more complex samples. This ability has come about in large part from both theoretical and experimental work associated with two-dimensional (2D) chromatography and the need to analyze such complex samples arising from metabolomic and proteomic studies. In the case of 2D-LC, there is a plethora of means by which two LC systems may be coupled. These include, but are not limited to: comprehensive versus "heart-cutting" techniques, columns in parallel or in series, the use of a different separation technique for each dimension (size exclusion chromatography followed by reverse phase, RPLC) or simply using the same separation technique but changing the chromatographic conditions (RPLC in both dimensions but with two very different columns, mobile phase conditions, *etc.*) [3]. All of the different data sets analyzed in this work are derived from the analysis of samples by comprehensive two-dimensional liquid chromatography with diode array detection (LC × LC-DAD) using reversed phase columns in both dimensions.

More complex samples imply more complex data to be analyzed. This has led to the increased need for chemometric methods capable of multiway data analysis. Chemometrics, as defined by Wold, aims to answer the questions "how to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such

information into data" [4]. To date however, processing of such data traditionally requires a decrease in the dimensionality of the data leading to a loss of information [5-7]. Chemometrics has been aimed at pattern recognition, quantification, classification and ranking capable of modeling the entire data set [5, 8]. Such methods include principal component analysis (PCA) [9], parallel factor analysis (PARAFAC) [10], multivariate curve resolution-alternating least squares (MCR-ALS) [11] and cluster analysis (CA) [5]. As with any analytical approach, the chosen chemometric technique must be capable of solving the analytical quandary at hand, and it must also be an appropriate technique for the data being analyzed. In this research, the overall objective was the resolution of chromatographic peaks (both from the background signal and from coeluting compounds) and quantification of said resolved compounds arising from LC × LC –DAD analysis of complex samples. Hence, chemometric methods aimed at curve resolution, MCR-ALS and PARAFAC are of interest. It is of the utmost importance that the data meet all of the requirements set forth by the algorithm in order to achieve a chemically meaningful result.

The background information required to enhance the readers' understanding of Chapters 5 through 8, that detail the research results from this project, is covered in Chapters 2 through 4. Chapter 2 reviews the theory and instrumentation of liquid chromatography, and it touches on why 2D separations are of importance. LC × LC is expanded upon with respect to the use of high temperatures, increased peak capacities and quantification. The theoretical background with respect to the chemometric methods used in this work is presented in Chapter 3 along with nomenclature and visual representations used to describe the four-way data analyzed here. Chapter 4 describes the different sample types analyzed (urine, wastewater treatment plant effluent and wine) in this work and their relevance to LC × LC separations. The details of the

analysis and the chromatography conditions for each of the three complex samples analyzed can be found in this chapter.

The development of the chemometric method and its application to replicate urine samples is described in Chapter 5 and in reference [1]. After a close inspection of the data, it was determined that the chemometric method would need to be able to handle large data sets, to be unaffected by retention time shifting in both chromatographic dimensions, and to be able to resolve rank deficient data, either chromatographic or spectral rank deficiency. This was achieved by the combination of iterative key set factor analysis followed by alternating least squares analysis utilizing the spectral selectivity constraint (IKSFA-ALS-ssel). A comparison of this approach to that of a PARAFAC-based approach previously described by the Rutan group [2] was undertaken, and the IKSFA-ALS method was determined to have several advantages over the PARAFAC approach. The most probable cause of disparities between the methods is that PARAFAC requires trilinear data, while IKSFA-ALS does not. Several of the parameters involved in this method require a subjective input by the user. A standards mixture analysis was used to investigate the effect such parameters have on quantification. In the section of chemometrically analyzed urine control data, over fifty peaks were found and of those thirty-four were resolved well enough for quantification. Precision of quantification was determined via percent relative standard deviation (% RSD) calculations which ranged from 0.09 to 16.

The urine control data and standards mixture data analyzed in Chapter 5 [1] did not consist of a calibration or standard addition set of samples, but rather of fourteen replicates; hence, only precision of the IKSFA-ALS-ssel method could be investigated. The goal of Chapter 6 was to determine both precision and accuracy of the chemometric method. That work is published in reference [12]. Resolution and quantification of only one target analyte

(phenytoin in wastewater treatment plant effluent) is the interest of this work as opposed to the goals of resolution and quantification of as many compounds as possible as was desired in the work discussed in Chapter 5. The wastewater treatment plant effluent (WWTPE) was extremely complex. To accomplish an acceptable resolution of the phenytoin peak from the interferent peak in a reasonable run time, it was deemed necessary to utilize three stages of chromatographic separations, thus the application of IKSFA-ALS for the first time to 3D-LC data. A sample selectivity constraint was also employed for the first time to correct the overfitting of the calibration samples and thus improve the chemometric resolution of the spiked phenytoin compound with the respect to the chromatographic interferent. The concentration of phenytoin in the WWTPE sample was determined to be 42 ± 1 ng/L, which is not significantly different from that of the 2D-LC/MS/MS reference method. It is interesting to note, that the precision of the IKSFA-ALS method applied to the LC × LC-DAD data was significantly better than that of the precision of the reference method.

To further both the fields of chemometric curve resolution and LC × LC chromatography, an investigation of the chromatographic factors that affect chemometric quantification was undertaken in reference [13] and is discussed in Chapter 7. To date very little research has been focused on the quantification of LC × LC data. Several different methods of peak quantification were compared (LCImage software, a total summation method and the manual baseline method) for both quantification of the raw data and the chemometrically resolved data. The manual baseline method was shown to yield better precision of quantification for both the raw and chemometrically resolved data. Chromatographic factors such as data complexity, retention time shifting, chromatographic and spectral peak overlap, large dynamic range and background signal interference were found to greatly influence the precision of quantification. Each of these

factors was thoroughly investigated and reported in reference [13] and Chapter 7 herein. The IKSFA-ALS-ssel resolved data exhibit a 2.5-fold increase in precision of quantification as compared to the quantification of raw data.

This dissertation therefore chronicles the importance of both LC × LC separations and of chemometric analysis of data arising from 2D and 3D-LC, the development of a chemometric method for both resolution and quantification and the factors that affect the precision and accuracy of the quantification. While Chapters 5 through 7 show that both good precision and accuracy are achievable (and yield better results than that of quantification without chemometric resolution), the drawback to the method, and many other available chemometric techniques, is the time required to achieve the desired results, along with the necessary skill of the analyst involved. Chapter 8 seeks to alleviate that issue to some extent [14]. The goal of many analyses of such complex mixtures is not the identification and quantification of every compound in the sample; but instead, the goal is to determine which compounds are significantly different between different sample types and the quantification of only those compounds. This is typically based on either absence or presence of specific compounds from one sample to the next, but more often on a significant concentration change of a compound between different samples. If an appropriate chemometric method could be found to determine which of the peaks in a complex sample exhibited significant concentration differences, all other peaks could be eliminated from further analysis such that only a few peaks of interest would be left for identification and quantification. To this end, two rapid screening methods were investigated, the similarity index (SI) method and the Fisher ratio (FR) method. Both experimental data (analysis of wine samples from three different vineyards) and simulated data were subjected to both methods. Several statistical analyses are also described that were used to verify the SI and

FR results and were also used to aid in the understanding of those results. Both chemometric methods were shown to be simple to use, to be rapid and to greatly reduce the number of peaks that would require further analysis. Through simulated data investigations, it was determined that the SI method was less affected by retention time shifting and by the background contribution. Also, the spectral information associated with every compound does not contribute to the SI value, unlike the FR method.

As the field of multi-dimensional chromatography continues to grow, gains popularity and applicability, and as the field of chemometrics advances with advances in computer technology, it is vital that chromatographers and chemometricians continue to collaborate. As chemometricians understand the needs of the chromatographers and the limitations of the chromatography itself, better algorithms can be designed that are faster, more accurate, and more automated for ease of use. As chromatographers understand the limitations of the current chemometric methods, chromatographic separations can be designed to minimize the chromatographic problems (such as retention time shifting) that reduce the usefulness or even completely prohibit chemometric analysis. The resulting dissertation is one very small step in the direction of collaborative work that may help lead to the realization of the full potential of each of these exciting fields of chemistry.

Chapter 2: Liquid Chromatography

2.1 Instrumentation

Liquid chromatography (LC) is a technique in which high pressure is utilized to force a liquid mobile phase and an injected sample through a column containing a bed of stationary phase consisting of micro-scale sized particles. Compounds present in the sample that have a chemical affinity for the stationary phase are retained longer on the column as opposed to compounds without or with a lesser affinity for the stationary phase; those compounds pass through the column more quickly. In this manner, compounds in a liquid sample can be separated in time. A schematic diagram, Figure 2.1, illustrates the major components associated with a high performance liquid chromatography instrument. The mobile phase in reversed phase liquid chromatography is a polar solvent (typically water, methanol or acetonitrile); while the stationary phase within the column is non polar (typically a hydrocarbon moiety bonded to silica particles). Either isocratic (constant delivery of a single solvent) or gradient (multiple solvents delivered such that the ratio of the differing mobile phase is changed either in a linear or stepwise manner) elution can be employed. In gradient elution, the polarity of the mobile phase composition is decreased so that the more nonpolar compounds will elute from the column over a reasonable time frame.

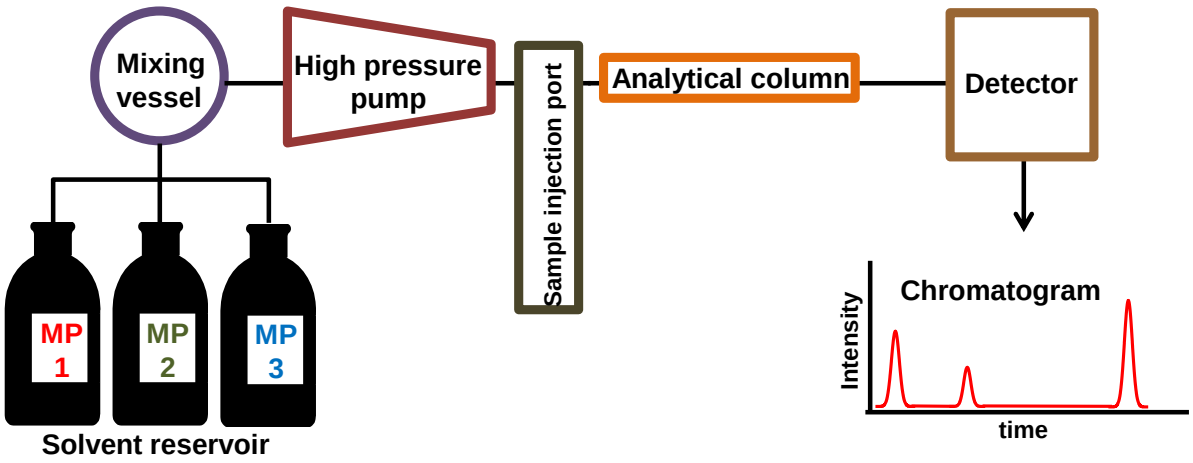

Figure 2.1: Schematic diagram of a liquid chromatograph. MP1, MP2 and MP3 are three different solutions that can be combined in the mixing vessel in differing ratios to be passed through the chromatographic instrumentation via a high pressure pump. The sample to be analyzed is then injected into the system such that the mixture compounds are separated while traveling through the analytical column. The results are displayed as a chromatogram that plots the detector response as a function of elution time.

While 1D-LC is an ideal method for the separation of many mixtures, the more complex the sample to be separated, the longer the required run time will be to separate the compounds and to completely elute the sample off the column, if well resolved peaks are the goal. For a complex sample, such as urine, Figure 2.2, a thirty minute sample run time is insufficient for the resolution of the compounds present in this complex mixture. Thus, there is the need for a method that is rapid and can yield good resolution for the hundreds of compounds associated with complex samples such as urine, wine and waste water treatment plant effluent. One such possibility is the use of two-dimensional liquid chromatography (2D-LC).

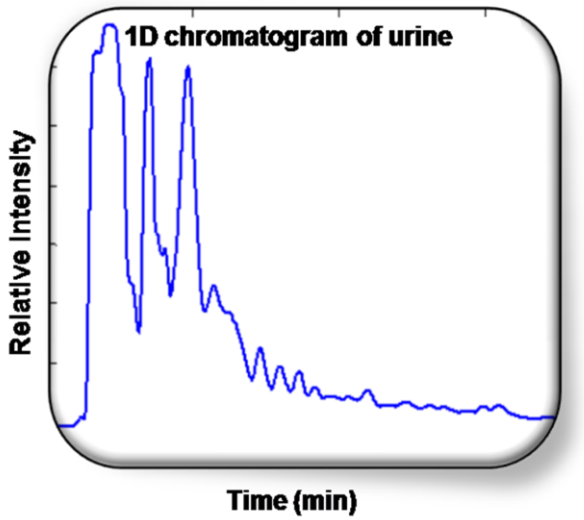

Figure 2.2: Chromatogram of urine sample resulting from a 30 minute 1D-LC separation. The chromatographic conditions are as follows: gradient elution from 0 to 70% B from 0 to 23 min, where A is 20 mM sodium phosphate, 0.1 mM EDTA at pH 6 and B is acetonitrile, with a flow rate of 0.1 mL/min. The stationary phase is a lab-made hydroxylated-hypercrosslinked material [15] packed in a 200 mm x 1.0 mm column.

2.2 Two-Dimensional Liquid Chromatography

Two-dimensional liquid chromatography is a technique in which the sample is passed through two independent column systems to achieve separation. This can be accomplished in one of two ways. In "heart-cutting" two-dimensional liquid chromatographic (LC-LC) methods, only a targeted portion of the separated first dimension (^{1}D) column effluent is transferred to the second dimension (^{2}D) column for further resolution. In comprehensive two-dimensional liquid chromatographic (LC × LC) methods, all of the effluent from the ^{1}D separation is sequentially introduced into the ^{2}D separation system to achieve a better resolution of overlapped peaks [3, 16]. In this way, the second dimension system (sampling device, column and detector) can be thought of as a chemically sensitive detector [17]. The chief advantage of LC × LC over 1D-LC is the potential for a greater resolving power [3, 16]. This can be shown by comparing the 1D-

LC chromatogram of urine in Figure 2.2 with the LC × LC separation of the same urine sample in Figure 2.3. The 1D-LC analysis does not offer sufficient chromatographic resolution of any of the compounds present in the sample, while there are over twenty-four chromatographically resolved peaks as can be seen from a simple visual inspection of the contour plot at a single wavelength of the LC × LC separation. The major drawback, until recently, has been excessively long run times, from an hour to a full day, required for a single sample injection due to the limited speed of the second dimension separation [2, 3, 16, 18, 19]. This disadvantage is

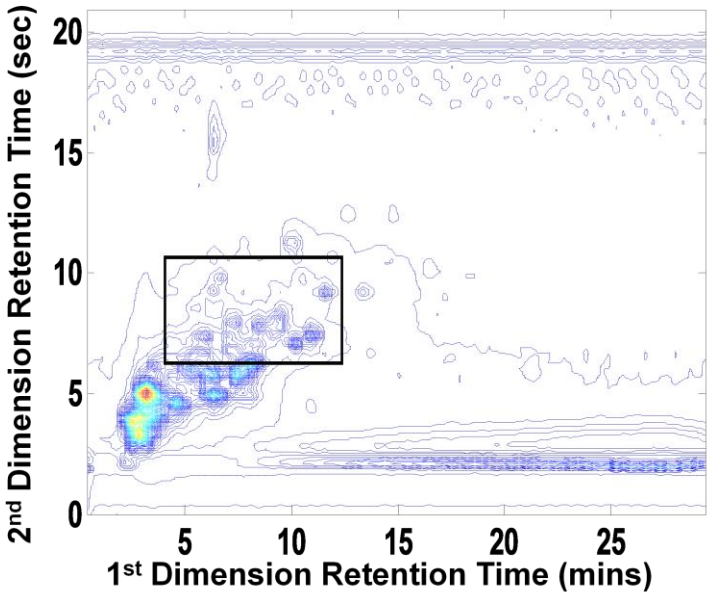

Figure 2.3: Contour plot of an LC × LC separation of the same urine sample shown in Figure 2.2 and the chromatographic conditions for the ^{1}D column are the same as provided in the caption of Figure 2.2. The chromatographic conditions for the ^{2}D column are as follows: gradient elution from 0 to 100% from 0 to 17.45 s, where A is 20 mM phosphoric acid and B is acetonitrile, with a flow rate of 3 mL/min. The re-equilibration time was 3 seconds. Two pumps are used to deliver the samples loaded in the two 35 µL loops in an alternating fashion to the ^{2}D column, where the ^{2}D stationary phase is a carbon-clad zirconia material packed in a 33 mm x 2.1 mm column [22].

being overcome by multiple different approaches. The use of monolithic columns [20, 21], ultra high pressure liquid chromatography (UHPLC) in the second dimension [23] and two ^{2}D columns run in parallel [24] have all been investigated. Stoll *et al*. [25] used a high temperature

gradient elution in the ^{2}D column to decrease the gradient cycle time of the second dimension to about 21 seconds for a corresponding 30 minute overall two-dimensional analysis time [25, 26]. Expanding on the use of high temperature, Stoll also introduced a method coined selective comprehensive multidimensional liquid chromatography (sLC × LC) in which "select" regions of the ^{1}D effluent are then analyzed in a comprehensive manner with the use of a ^{2}D column [27].

2.3 High Temperature in Fast LC × LC

As briefly discussed, Stoll *et al.* have overcome the extended run times required for LC × LC analysis with the use of high temperature in the ^{2}D separation [3, 25]. The LC × LC system illustrated in Figure 2.4 is a dual gradient system employing high temperature in the second dimension through the use of an eluent preheater and a heating jacket placed around the ^{2}D reverse-phased column. The sample is injected onto the first dimension system in which the eluent is preheated to 40 ^{o}C before passing through the ^{1}D column. A 10-port valve captures the effluent exiting the first column in either loop #1 or loop #2. As one loop is filling, the effluent captured in the other loop is injected onto the second column using a binary pump. In this manner all of the effluent from the first column is sequentially injected onto the second column for further separation of compounds. The second column employs temperatures greater than 100 ^{o} C. In this example, diode array detection (DAD) was employed after the second separation column. By reducing the second dimension run time (a limiting factor in the total analysis time) to a mere 21 seconds through the implementation of a high temperature second dimension system, the overall injection run time is significantly reduced.

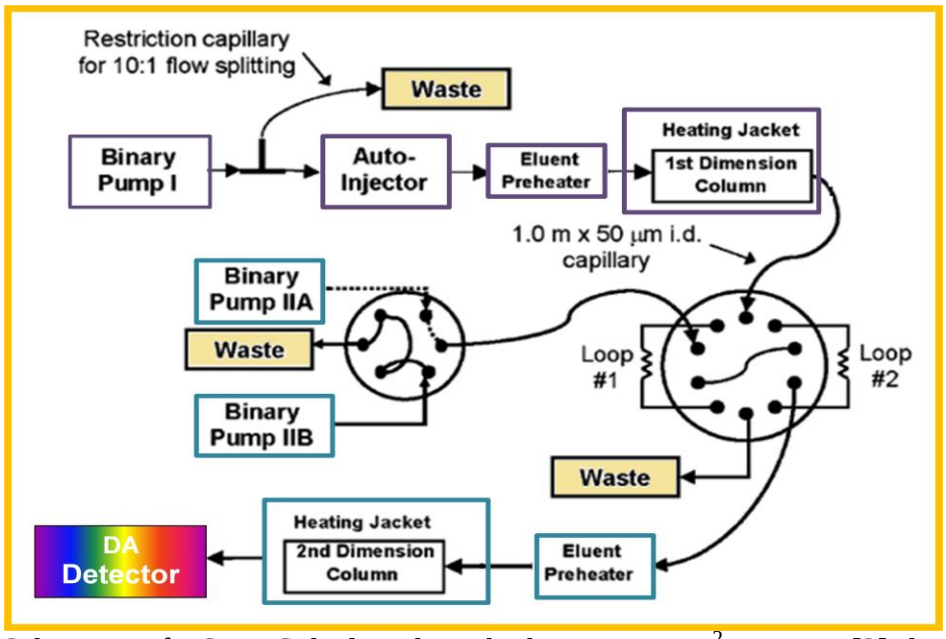

Figure 2.4: Schematic of LC × LC dual gradient, high temperature ^{2}D system [3] showing the use of three binary pumps, two eluent preheaters, heating jackets around both the ^{1}D and ^{2}D dimension columns, a six port switching valve for the ^{2}D gradient mobile phase and a ten port switching value that collects and delivers the ^{1}D aliquots to the ^{2}D system.

The effect temperature has on the linear velocity of the mobile phase (flow rate) can be explained using the van Deemter equation, which relates plate height (a quantitative measure of column efficiency) to the flow rate of the mobile phase as follows:

$$H = A + B/u + (C_S + C_M)u \qquad (2.1)$$

where H is the plate height in units of cm, the coefficient A is the eddy diffusion term with units of cm; B is the longitudinal diffusion term with units of cm^2/sec; $C_S + C_M$ is the resistance to mass transfer between the stationary and mobile phases term with units of sec, and *u* (cm/s) is the linear velocity of the mobile phase. The narrower the chromatographic peaks, the smaller the H, thereby leading to a better separation. Figure 2.5 illustrates the relationship of plate height to the linear velocity of the mobile phase. While there is currently some debate over the nature of the equations necessary to describe the A, B and C constants in the van Deemter equation, what is

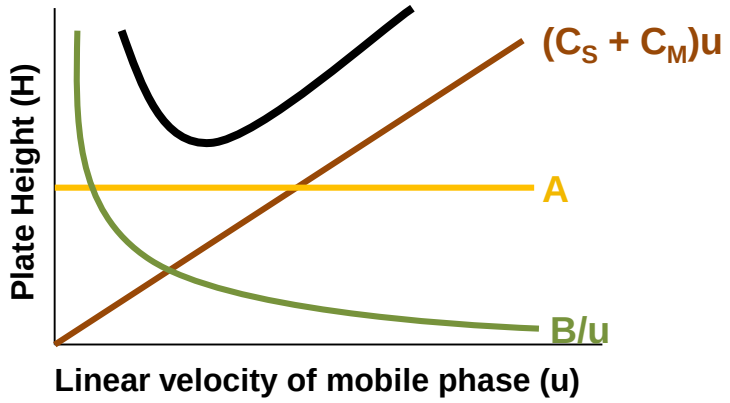

Figure 2.5: van Deemter plot for a packed LC column showing the relationship between the plate height (H) of the column and the flow rate of the chromatographic system where the black line represents the total plate height and the brown, yellow and green lines are the contributions due to the three different rate terms.

well established is the relationship of the B, C_S and C_M terms with the diffusion coefficient of the solute in the mobile phase, D_M. The longitudinal diffusion term, B, is directly proportional to the diffusion coefficient, D_M, and both mass transfer terms, C_S and C_M, are inversely proportional to D_M such that

$$H = A' d_p + B' \frac{D_M}{u} + \frac{C' d_p^2}{D_M} u \qquad (2.2)$$

where d_p is the particle diameter. This is significant because the diffusion coefficient for a liquid is temperature dependant as shown by Wilke-Chang equation [28, 29]

$$D_M = \frac{G\sqrt{\psi M_2}}{\eta V^{0.6}} T \qquad (2.3)$$

where T is the absolute temperature, η is the viscosity as a function of temperature, G is a constant, Ψ is the association constant for the solvent (1.0 for unassociated, non-polar, solvents, and 2.6 for water), M_2 is the molecular weight of the solvent and V is the molar volume of the

solute. A clear relationship can now be seen between the diffusion coefficient, temperature and eluent viscosity, which is also temperature dependent [3, 28, 29]. Intuitively, as the temperature of a liquid increases, its viscosity decreases; and thereby, the diffusion coefficient is increased. This has a direct effect on the longitudinal diffusion term (B) and the mass transfer terms (C) of the van Deemter equation, since B is proportional to D_M and C is inversely proportional to D_M, as seen from equation 2.2. Hence, as temperature increases, the viscosity decreases and the diffusion coefficient increases such that the C term in the van Deemter equation decreases and the B term increases. The effect on the van Deemter plot is illustrated in Figure 2.6. At higher temperatures, the van Deemter curve flattens out and the minimum plate height, H_{min}, is shifted to the right corresponding to faster flow rates. It is important to note that an increase in temperature, in and of itself, does not improve the efficiency of the column, *i.e.,* the H_{min} remains unchanged when all other parameters remain the same [29]. This concept is better understood by considering the equations that result from taking the derivative of H with respect to u. This leads to the optimal linear velocity equation (u_{opt})

$$u_{opt} = \sqrt{\frac{B}{C}} \tag{2.4}$$

and the minimal plate height (H_{min}) at this velocity is

$$H_{min} = A + 2\sqrt{BC} \tag{2.5}$$

Since B α D_M and C α $1/D_M$, the optimal velocity is directly proportional to the diffusion coefficient and the minimal plate height is proportional only to the A term of the van Deemter equation. However, the decrease in viscosity allows for smaller particle sizes to be utilized, which decreases the plate height and increases column efficiency.

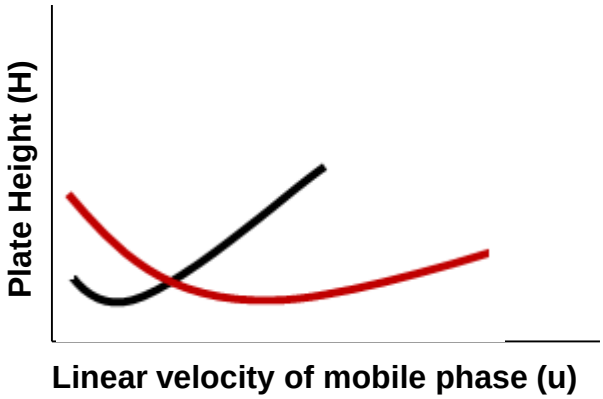

Figure 2.6: Illustration of two van Deemter curves at different temperatures. The black curve represents the use of a low temperature, while the red curve represents the system at a higher temperature. Note that the value of H_{min} is unchanged for both curves, but now corresponds to a higher linear velocity for the high temperature curve.

2.4 Peak Capacity

A significant advantage of LC × LC separations is the enhanced resolving power as compared to that of 1D separation methods [30]. O'Farrell in 1975 [31] and Erni and Frei in 1978 [32] showed that for complex samples, 1D separations are incapable of providing sufficient selectivity or peak capacity [16]. The statistical overlap theory (SOT) by Davis and Giddings also showed that the possible peak capacities achievable by 1D-LC are not sufficient for the resolution of complex, multi-constituent samples [30]. Using SOT for 1D-LC, it was shown that samples consisting of only 10-20 components would generate chromatograms where multiple, seemingly single component peaks, actually consist of two or more overlapping, unresolved components. It is therefore apparent that exceptionally high peak capacities are required for the separation of complex samples containing hundreds of components [3]. A brief review of the concepts of chromatographic resolution and peak capacity will aid in the understanding of the issues associated with maximizing the peak capacity of a 2D separation system.

Resolution, R_s, is a measure of the ability of the chromatographic system to resolve two adjacent analyte peaks and is given by:

$$R_s = \frac{t_2 - t_1}{(w_2 + w_1)/2} \tag{2.6}$$

where t_2 and t_1 are the peak retention times of two adjacent peaks, and w_2 and w_1 are the corresponding peak widths at the peak base. Baseline resolution of the two adjacent peaks is achieved when $R_s = 1.5$, and sufficient resolution is typically considered to occur when $R_s = 1.0$. If the resolution is too low, the peaks will be overlapped (unresolved); on the other hand, a large resolution may lead to a separation in which the peaks are very far apart, unnecessarily increasing the analysis run time. The peak capacity, n_c, is another quantitative measure of the quality of the separation of the chromatographic system. Peak capacity, unlike resolution which takes into account only an adjacent peak pair, looks to give a measure of chromatographic separation of the entire available chromatographic "space"; *i.e.*, how many peaks with a given resolution can be observed within a given time interval. The peak capacity (n_c) for a 1D gradient system is defined as:

$$n_c = 1 + \frac{t_{R,last} - t_{R,first}}{w} \tag{2.7}$$

where $t_{R,last}$ is the retention time of the last peak to elute from the column and $t_{R,first}$ is the retention time of the first peak to elute from the column. The width (w) is the base width as before and is typically approximately constant for gradient methods. This equation assumes a resolution of one. It is tempting to then use the multiplicative rule and state that the peak capacity of a 2D system ($n_{c,2D}$) is the product of the peak capacity of the first dimension of a 2D separation (^1n_c) and the peak capacity of the second dimension of a 2D separation (^2n_c) such that

$$n_{c,2D} = {}^{1}n_c \times {}^{2}n_c \quad (2.8)$$

This definition for two-dimensional peak capacity implies that the system has achieved the impractical and improbable "ideal" chromatographic conditions, and hence typically over estimates the systems separation power [3, 33, 34]. According to Giddings, [30] the following criteria affect a systems ability to reach this "ideal" state [35]. First, the two separation systems used for the 2D analysis must be orthogonal and completely independent. There can be no correlation between the analyte retention of the two columns used for separation. Orthogonality is achieved when both phase systems are completely uncorrelated allowing for the total use of the separation space. It is worth noting that both analytical methods must be appropriate techniques for the separation of the analytes in question, thus limiting the possible combinations of techniques that can be employed.

A lack of orthogonality directly impacts Giddings' second criterion which requires that the entire available separation space must be utilized by the sample constituents. If the mechanisms of separation are totally correlated, a separation diagonal results as illustrated in Figure 2.7 A. When orthogonality is achieved, the entire separation space is occupied (Figure 2.7 C); while moderate correlation can yield partial coverage of the separation space (Figure 2.7 B). Chromatographic conditions, such as mobile phases and flow rates, must be compatible between the two dimensions; because of this, it is sometimes necessary to use less than optimal chromatographic conditions to achieve compatibility between the two systems [35]. This compromise also has the deleterious effect of reducing the available separation space. It is, therefore, important to have a metric, a correction factor, which is capable of defining the useful retention space. Rutan *et al.* [36] compared several geometric methods to ascertain their applicability in the determination of what was termed fractional coverage ($f_{coverage}$); *i.e.*, "the

fraction of the separation space (area) that can in principle be covered by peaks". The authors were interested in determining a $f_{coverage}$ metric based on the *separation* dimensionality of the chromatographic system (the fraction of the separation space that *can be filled* by peaks) and is independent of the *sample* dimensionality (the fraction of the available space that *is actually filled* due to the separation of the sample constituents). In this manner the lack of diversity of the sample itself will have no bearing on the value of the calculated $f_{coverage.}$ This concept is illustrated in Figure 2.7 such that the area inside the red lines for each of the depicted

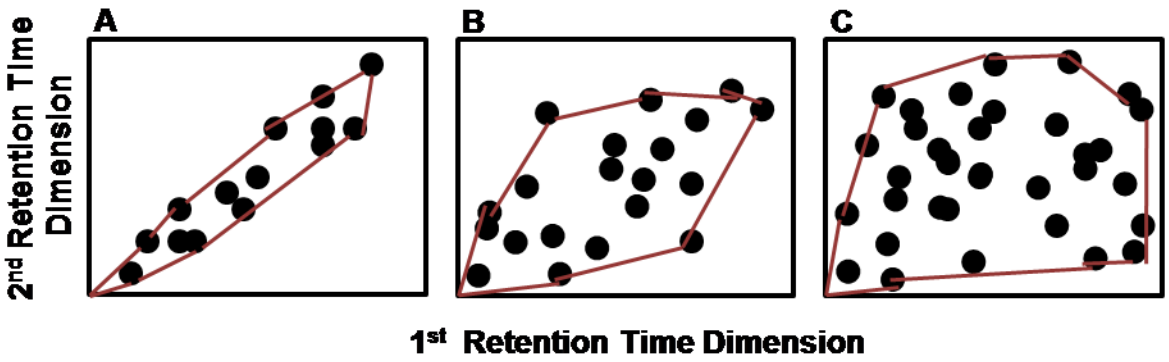

Figure 2.7: Illustration of peak coverage of the available separation space for a 2D separation system under the conditions of (A) non-orthogonality (B) partial orthogonality and (C) complete orthogonality. The red lines illustrate the concept of $f_{coverage}$ if the minimum convex hull method was applied to the illustrated data.

orthogonality conditions is representative of the $f_{coverage}$ metric. Rutan *et al.* concluded that the minimum convex hull method for the determination of $f_{coverage}$ fulfilled all of the criteria set forth by the authors and offered several advantages over the other methods investigated, including simplicity and ease of use. This correction factor can now be included in the peak capacity equation 2.8 to correct for the lack of orthogonality of the separation.

$$n^{*}_{c,2D} = {}^{1}n_c \times {}^{2}n_c \times f_{coverage} \qquad (2.9)$$

The last criterion set for by Giddings that must be met in order for equation 2.6 to be accurate is that the separation achieved from the first separation system must not be lost upon implementation of the second separation system. While 2D separations produce order of magnitude greater peak capacities than those of 1D-LC separations, implementation of the analysis in a timely fashion (fast LC × LC) decreases the resolving power of the technique along the first dimension separation [3, 37]. Sampling-induced first dimension peak broadening occurs due to remixing that occurs in the sampling device of the first dimension effluent while awaiting injection onto the second dimension column. Since both resolution and peak capacity are inversely dependent on peak width, a decrease in both resolution and peak capacity of the first dimension separation occurs. Murphy, Schure and Foley, (MSF) working toward the realization of ideal peak capacities, centered their work on efficient sampling of the first dimension effluent in order to maximize the first dimension resolution [34]. A significant consequence of their work is the M-S-F sampling criterion which states that for an 8σ (where σ is the standard deviation of the peak) first dimension peak width, the effluent must be sampled at least three to four times to avoid first dimension resolution loss [3]. They concluded that, "the shortest sampling time in to the second dimension gives the best resolution and the longer sampling times decrease resolution along the first dimension axis."

Hence another correction factor to equation 2.6 (the ideal peak capacity of a two-dimensional separation) is needed to account for this peak broadening effect. Using Statistical Overlap Theory (SOT) to predict the number of observed peaks in a 2D simulated data set, Davis *et al.* [38] showed that the average first dimension peak broadening factor, $<\beta>$, can be calculated from

$$<\beta>=\sqrt{1+\kappa(t_s / {}^1\sigma)^2} \qquad (2.10)$$

where κ is a fitting coefficient (equal to 0.214 for Davis predictions of <β>), t_s is the sampling interval and $^1\sigma$ is the standard deviation of the ^{1}D peak before sampling. By taking into account the two correction factors, $f_{coverage}$ and <β>, the effective two-dimensional peak capacity ($n^*_{c,2D}$) can be calculated as follows:

$$n^*_{c,2D} = {}^1n_c \times {}^2n_c \times \frac{1}{<\beta>} \times f_{\text{coverage}} \tag{2.11}$$

It is directly relevant to this work, and therefore necessary, to briefly mention at this point that the loss of resolution can be effectively overcome, at least in part, by the application of appropriate chemometric methods (to be discussed in Chapters 3, 5-7) that mathematically resolve and quantify overlapped peaks; thereby, essentially enhancing the resolving power of the 2D separation and reducing the conundrum faced by many chromatographers forced to choose between increased resolution or decreased run times [37]. Davis *et al.* showed that the minimum resolution required to observe two peak maxima in a simulated 2D separation was reduced from 0.5 without the assistance of chemometrics to 0.256 for chemometrically assisted resolution and is even further reduced with the addition of spectral information acquired from diode array detection (DAD) [39].

2.5 Possible Detectors

The most commonly used detectors in LC × LC are diode array detectors (DAD) and electrospray ionization (ESI) mass spectrometers (MS) [3]. It is important to note that the chosen detector must be capable of very high scan rates. By the very nature of these detection methods, there are two obvious limitations. In the case of DAD, only light absorbing compounds will be detected; while the compounds of interest must be charged for detection by MS. As usual

with any analytical method there are advantages and disadvantages associated with either of the above detection methods. ESI-MS can be used in the field of proteomics, for the identification of peptides and thus proteins making MS detection vital. There are however two main impediments with respect to MS detectors when coupling them to LC × LC instrumentation: accuracy of quantification is affected by ion suppression; mobile phase choices are limited; and the slow m/z scanning speed of some mass analyzers is inadequate [3, 40]. Diode array detection is capable of 100 scans/s, which is more than sufficient for fast LC × LC. The most significant advantage, especially when analyzing complex samples such as biofluids, is the precision of the detection achieved with DAD leading to high precision in quantification [2, 3].

2.6 Peak Quantification

While the reduction in run times and the increase in peak capacities make LC × LC an ideal technique for the analysis of complex samples, such as urine, wine and wastewater treatment plant effluent as described in Chapters 4-8, it is of the utmost importance that precise and accurate quantification of those compounds be achieved also. Such complex samples may also arise from proteomic and metabolomic studies. In many instances, the goal of such studies is identification (by changes in concentration or in concentration ratios) of potential biomarkers; thus, the ability to accurately quantify both major and minor constituents in a sample is of great significance in proteomics and metabolomics [41]. Unfortunately, work on quantification in LC × LC is exceedingly sparse; only a handful of reports are available. Table 2.1 summarizes research to date in the literature of LC × LC quantification [42-46].

Table 2.1: Methods used for the quantitative analysis of LC × LC data.

Authors	Year	Method	Compounds	% RSD
J. Pól, B. Hohnova *et al.* [43]	2006	Summation of 2nd dimension chromatograms	Acidic compounds in atmospheric aerosols	8 %
M. Kivilompolo T. Hyötyläinen [44]	2007	Summation of 2nd dimension chromatograms	Antioxidant phenolic acids	2 - 14 %
M. Kivilompolo V. Oburka *et al.* [45]	2008	Peak volume determination in 2D contour plots	Antioxidant phenolic acids	3 - 13 %
L. Mondello, M. Herrero *et al.* [46]	2008	Area summation of "data point triangles"	Aurapten/ coumarin Coumarin internal standard	0.1 - 3.0 % 5 %

In 2006 and 2007, the Hyötyläinen [43, 44] group obtained quantitative results from a LC × LC analysis based on the summation of the areas of second dimension chromatograms of several consecutive modulation periods (second dimension "slices"). In 2008, instead of using a summation of the second dimension chromatograms for quantification, Kivilompolo *et al*. [45, 47] applied software previously used in the quantification of comprehensive two dimensional gas chromatography (GC × GC) data. This method of quantification involves the determination of peak volumes in two-dimensional contour plots. Peak heights for each peak data point are calculated from the contour plots and multiplied by the area under the corresponding point. This method should be exactly proportional to the summation of the second dimension chromatograms method, as long as the same peak boundaries are used. Both of these methods can be considered as resulting in measured quantities.

Mondello *et al.* [46] developed an automated method for quantification in which the area under each second dimension peak is determined by summing areas of what they termed "data

point triangles". This method was employed in the analysis of eighteen calibration curves of aurapten in grapefruit essential oil with coumarin used as an internal standard. The authors report percent relative standard deviation (% RSD) values ranging from 0.1 to 3.0 for the aurapten/coumarin ratio and 5 for the coumarin internal standard. However, they do not fully explain how the algorithm determines peak baselines or how it handles non-ideal peak behavior such as phase shifting, tailing peaks or "embedded" peaks. Also, Reichenbach showed the mathematical equivalence (both graphically and with an equation) of the summation of the "data point triangles" method and the summation of the ^{2}D peak areas method [48].

Reichenbach *et al.* [22] developed automated software that uses the "watershed" algorithm to determine a two dimensional peak boundary either before or after the background is subtracted from the total signal; all data points within this area are then summed. While this method is quite convenient, it does not take into account the chromatographic nature of a two-dimensional peak. Vivó-Truyols *et al.* [49] recently published results of a study on the use of the watershed algorithm for the detection of comprehensive two-dimensional chromatography peaks and describes two drawbacks to the algorithm. The first is that the watershed algorithm "does not impose the condition of continuity for a peak". This issue is shown for LC × LC data in Figure 2.8. Figure 2.8A is the two dimensional contour plot of an LC × LC peak such that the red lines are the boundaries of that peak as determined by LCImage software. The value of every data point within that discontinuous box (from slices 2-10) is summed to give the volume of the peak. However, from Figure 2.8 B (the sequenced second dimension chromatogram plot) peak 1 appears to consist only in slices 4-7. If slice 5 and slice 10 (Figure 2.8 C and D respectively) are plotted separately, it is apparent that slice 5 is integrated in a continuous manner while slice 10 is not continuous, e.g., this implies that there are 4 chromatographic peaks

associated with the same compound within a given second dimension slice. The second drawback cited by Vivó-Truyols is that any second dimension retention time shifts must be corrected prior to using this algorithm.

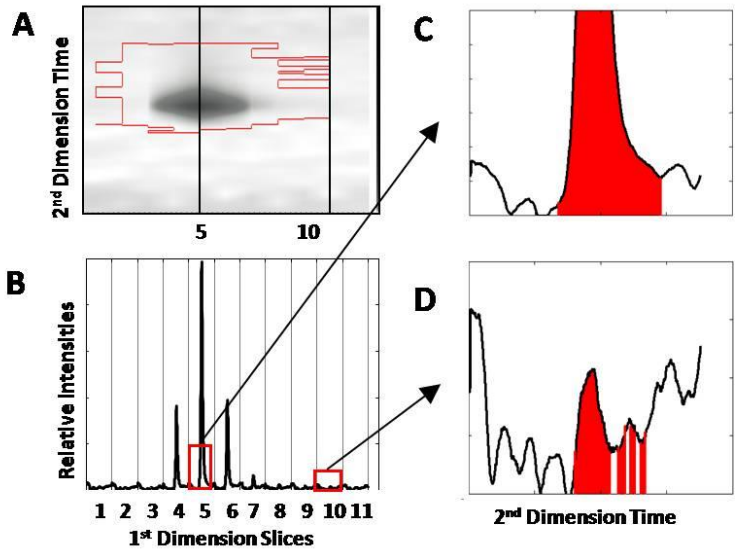

Figure 2.8: (A) Contour plot of a LC × LC peak after background subtraction using LCImage software. The (red) line is the peak boundary as determined by the LCImage software via the watershed algorithm. (B) The corresponding sequence of 2nd dimension chromatograms. (C) The integrated area of slice 5. (D) The integrated area of slice 10. (This peak corresponds to the standards mixture peak 1 as discussed in Chapter 4.)[13]

Thekkudan and Rutan [50] recently performed simulation experiments to determine the effects of retention time shifts and sampling period (modulation cycle) changes on LC × LC peak quantification. This was accomplished by varying the retention times and peak widths of simulated data. Peak quantification was determined by two methods, the moments method, in which the second dimension peak areas were obtained and then summed to yield the LC × LC peak volume, and the Gaussian fitting method, in which the consecutive second dimension peak areas were fit to a Gaussian model of a first dimension peak and the area of the first dimension Gaussian peak so obtained was used as the peak volume. The moments method was relatively

unaffected by retention time shifts or differences in peak widths yielding % RSDs consistently between 1.7 and 1.9 for their assumed signal to noise value. However, the Gaussian method yields % RSD values between 1.2 and 1.3 irrespective of retention time shifting as long as the simulated peak width was sufficiently wide to consist of at least three second dimension fractions [50]. These results for simulated data indicate the possibility of achieving reproducible LC × LC quantification, under ideal conditions, similar to that achieved by 1D-LC. The area summation method described by Thekkudan *et al*. [50] is equivalent to the manual baseline method (utilized in this research and discussed specifically in Chapters 5-8), and is based on the premise that the sum of the second dimension peak areas (slices) is equal to the volume of that LC × LC peak [48, 50].

In the vast majority of the reports, only well-resolved peaks were quantified. In many instances, however, the data arising from LC × LC analysis of complex samples will consist of multiple compounds that elute at very similar retention times and of multiple compounds that have the same or very similar spectra. This reality greatly limits the ability to achieve chromatographically anything resembling "ideal conditions" for the quantification of large data sets. It is therefore essential to employ a means of chemometrically resolving the overlapped peaks. There are several methods by which chemometrics can aid in our need for better resolution to achieve accurate and precise quantification of complex samples. Several of the methods that are used in this work are described in the following chapter.

Chapter 3: Chemometric Techniques and Theory

To aid the reader, a short discussion concerning the manner in which the terms are used in this work may be helpful. Up to this point, the discussion has focused on the chromatography side of the presented research in which the terms analyte, peak and compound are used somewhat interchangeably. At this point we change gears to focus on the chemometrics background that is necessary for further discussion. It is important to keep in mind that the terms *component* and *compound* are not synonymous. Many of the algorithms discussed in the subsequent sections and chapters look for the most different spectra, or components, within the data, as shown in Figure 3.1. Therefore, a four component model will find the four most unique spectra and each spectra will be assigned to its own component, *i.e.*, component 1 (purple), component 2 (teal), component 3 (red) and component 4 (yellow). Every chromatographic peak, compound, is associated with a specific spectrum and thus that *compound* is assigned to its corresponding *component.* Because of this, more than one compound (chromatographic peak) can be assigned to any given component if they are characterized by the same spectrum.

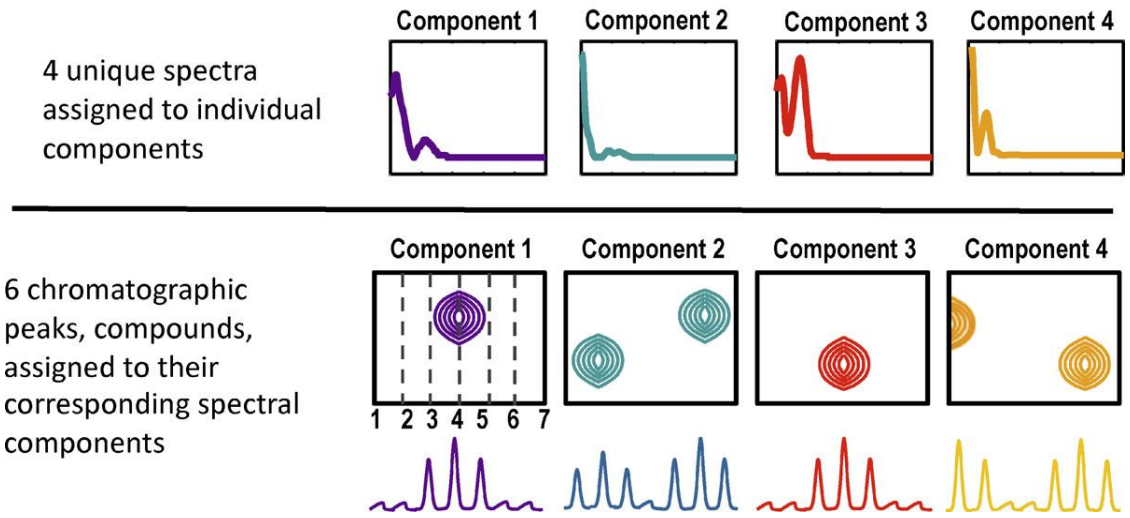

Figure 3.1: Illustration describing the relationship between components and compounds.

3.1 Data Structure

In the simplest case, the data from a single chromatographic experiment (1D-LC-DAD, which gives rise to two-way data) can be contained in a matrix **X**, which consists of absorbance values as a function of elution time and wavelength, and can be represented as follows:

$$\mathbf{X} = \mathbf{R} \cdot \mathbf{S}^{\mathbf{T}} + \mathbf{E} \tag{3.1}$$

where **R** is the chromatographic matrix, **S** is the spectral matrix and **E** is an error matrix [51]. The columns of the data matrix **X** are absorbance measurements that vary with time (chromatograms) and the rows are intensity measurements that vary with wavelength (spectra). The columns of matrix **R** contain the chromatograms of the individual pure components present in the sample represented by matrix **X**, while the columns of matrix **S** contain the spectra of those components, as depicted in Figure 3.2. In this work, the LC × LC- DAD data is collected by the instrument as two-way data for each separate 1st dimension injection such that all of the 2nd dimension injection chromatograms are sequenced end to end as seen in Figure 3.3A. In this

manner the data is represented as a two-way data matrix, X, with dimensions $IJ \times L$. Here, I is the number of data points in each 2nd dimension chromatogram, J is the number of data points in each 1st dimension chromatogram and L is the number of points in each spectrum. The data can be rearranged with dimensions $I \times J \times L$ (shown at one wavelength in Figure 3.3B).

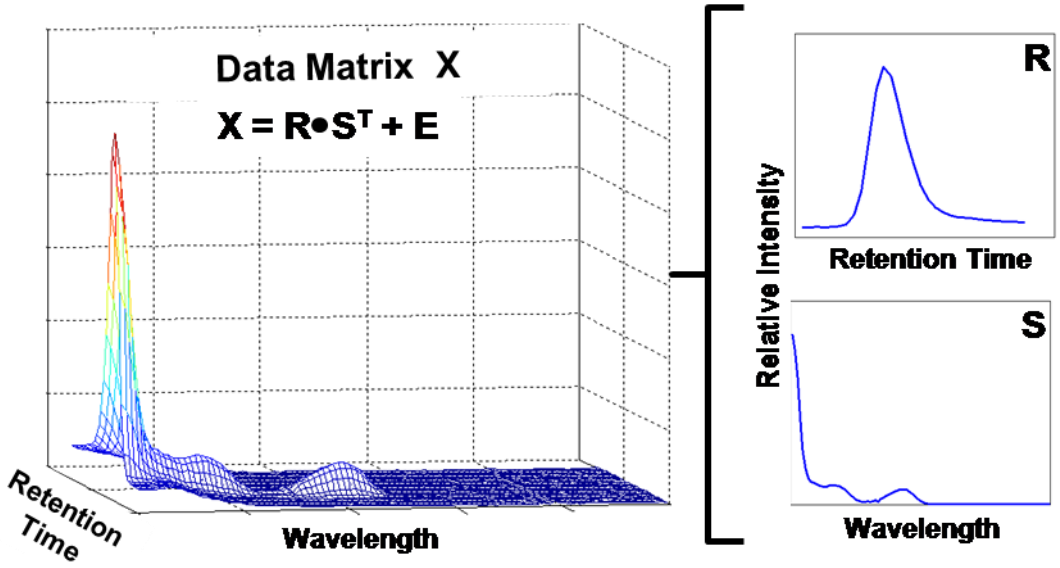

Figure 3.2: Mesh plot of a 1D chromatographic peak and corresponding spectrum and two individual plots of the corresponding **R** matrix (chromatogram) and **S** matrix (spectrum).

Four-way data generated by LC × LC-DAD analysis can be visualized as shown in Figure 3.4 with the data existing in multiple cubes. The rows and the columns are the first and second chromatographic dimensions, the slices of each cube contain the spectral dimension and each entire cube, as a whole, is an individual injection such that K is the number of different samples that were analyzed. In this way the four-way data array now has dimensions of $I \times J \times K \times L$. Four-way data can be unfolded into three-way or two-way data by reshaping the data. Figure 3.5 visually illustrates the reshaping of four-way data to three-way data by combining the individual injections, each data cube. The data is further unfolded by reshaping the two chromatographic

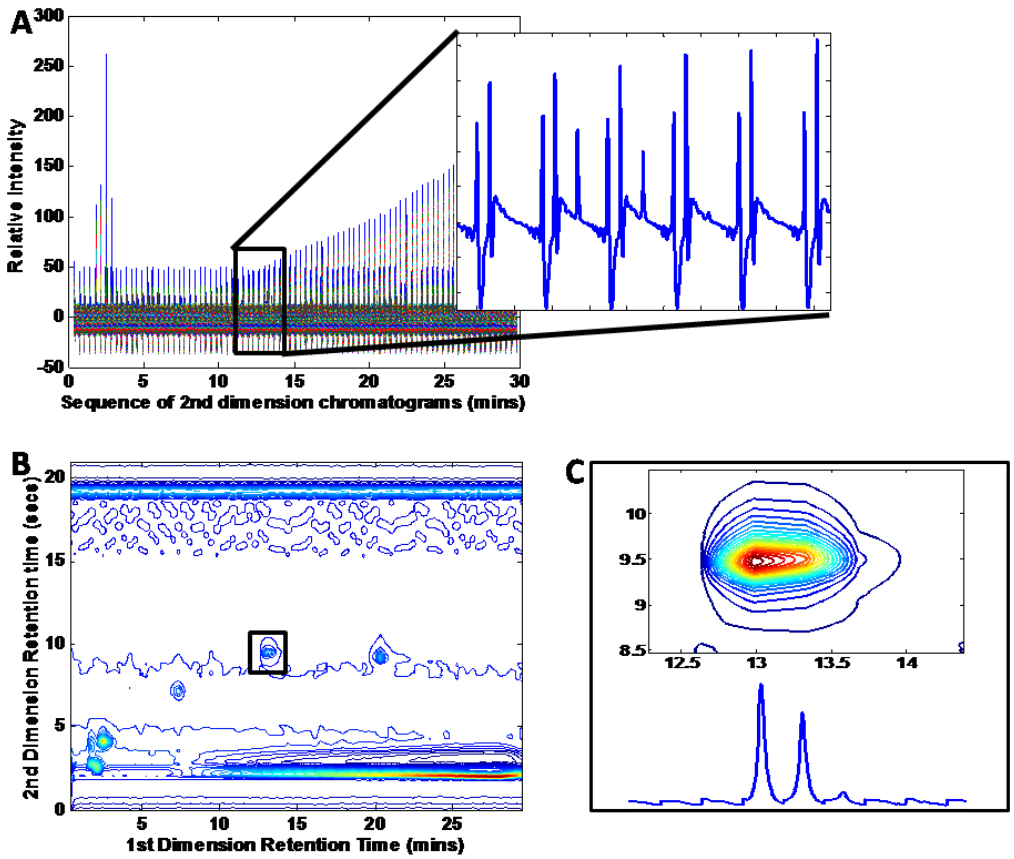

Figure 3.3: Several different types of plots for the 2nd sample of the standards mixture raw data. (See Chapter 4) (A) Plot of the raw data as collected by the instrument at 216 nm, such that all 2nd dimension injections of a corresponding sample injection are sequenced end to end. The insert shows an enlarged section of the sequenced data that was determined to encompass the corresponding peak illustrated in the contour plot at 216 nm of (B). Using the determined section dimensions, the same peak is shown as a contour plot at 216 nm in the box (C) along with its corresponding sequenced 2nd dimension chromatograms.

dimensions preserving the spectral dimension such that this dimension is not reshaped with any of the other dimensions of the data. It is important to realize that this procedure is not restricted to preserving the spectral dimension, $IJK \times L$. It is possible to preserve any of the four dimensions of the data such that the two-way data dimensions of $IJL \times K$ (preservation of the sample dimension), $ILK \times J$ (preservation of the first retention time dimension) and $KJL \times I$ (preservation of the second retention time dimension) are each possible arrangements of the unfolded data.

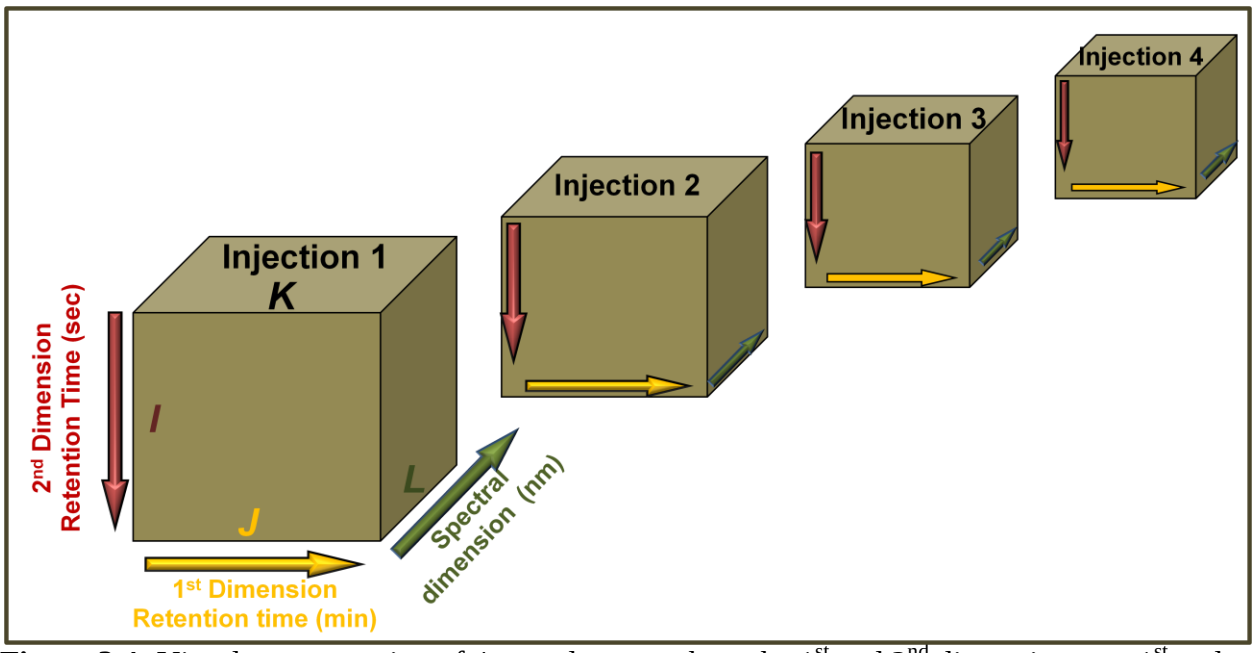

Figure 3.4: Visual representation of 4-way data set where the 1st and 2nd dimensions are 1st and 2nd chromatographic retention times respectively, the 3rd dimension is spectral wavelengths and the 4th dimension is the number of injected samples.

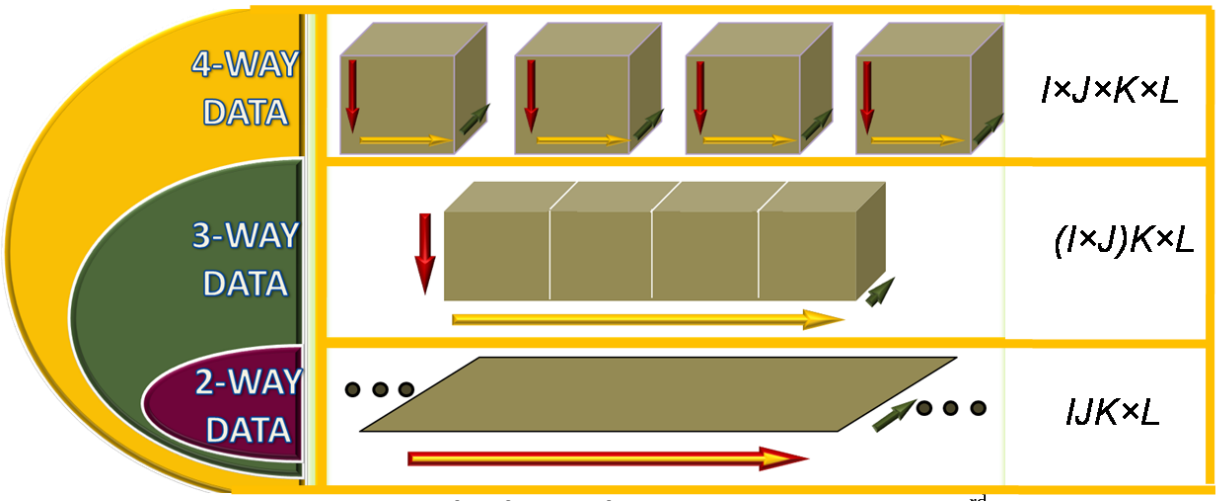

Figure 3.5: Visual representation of unfolding four-way data in which the 4rd mode (*i.e.*, the spectral dimension) is conserved during the unfolding of the four-way data to two-way data.

3.2 Singular Value Decomposition (SVD)

Many of the methods utilized for chemometric data analysis use singular value decomposition (SVD) as an initial step to determine the rank of the data, *i.e.*, the number of

components comprising the data set, *N*. This is due to the robustness and wide applicability of the algorithm [52, 53]. SVD is an eigenanalysis method that decomposes the data matrix **X** (with dimensions $M \times L$ such that $M=IJK$ in all data analyses performed in this work) into three matrices **U**, **W** and **V** and is represented as follows:

$$\mathbf{X} = \mathbf{U}\,\mathbf{W}\,\mathbf{V}^{\mathbf{T}} \tag{3.2}$$

where the columns of $\mathbf{U}(M \times L)$ contain the left singular vectors representing the variation in the rows of the data matrix. The columns of $\mathbf{V}(L \times L)$ contain the right singular vectors representing the variations in the columns of the data matrix. $\mathbf{W}(L \times L)$ is a diagonal matrix such that the elements on the diagonal are non-negative numbers representing the variance contribution of each principal component, *i.e.*, the singular values of the original data matrix **X**. This allows for a hierarchical ranking of each component's ability to explain the variation in the data. Most of the variance of the data is modeled by the first component, while each subsequent component models the maximum variance not described by the previous component. The greater the singular value, the greater the significance of that component; while typically, smaller singular values represent less significant contributions, such as noise, in the data [52]. In this manner, the number of significant components contributing to a data matrix can be determined using a scree plot, as shown in Figure 3.6. The scree test is based on the observation that the residual variance should level off when the remaining components account for only random error; thus, factors to the left of the "elbow" in the scree plot are retained. Truncation of the data (where *N*, the number of significant components) results in the following dimensionality changes $\mathbf{U}(M \times N)$ $\mathbf{V}(N \times N)$ and $\mathbf{W}(N \times N)$. The scree plot test is further discussed in Chapter 5.

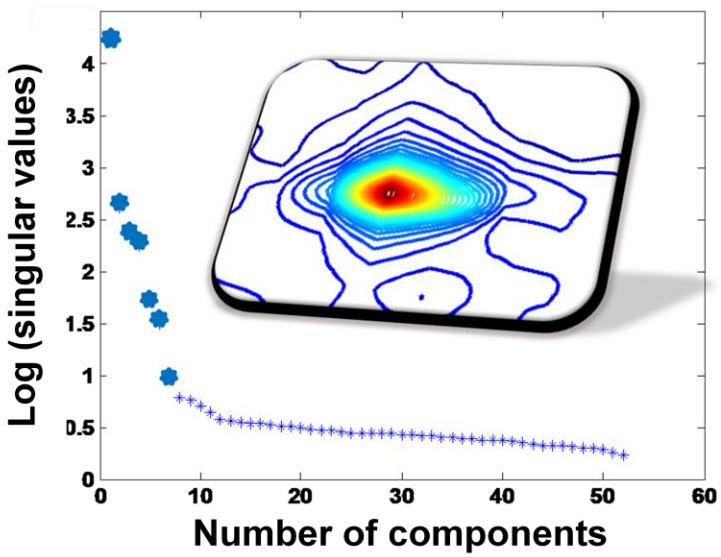

Figure 3.6: Scree plot showing the relative importance of the singular values plotted as a function of the number of factors in the data matrix.

3.3 Iterative Key Set Factor Analysis (IKSFA)

The IKSFA method was developed by Malinowski and is a preferred set selection method that assumes the purest spectra in the data set are mutually more dissimilar than the mixture spectra [54]. IKSFA is an iterative improvement over KSFA which seeks to find the minimum number (N) of spectra (out of M total spectra) required to represent the entire data set through the characterization of the most orthogonal spectra that typify the original data matrix [53, 55]. There are over ten billion possible combinations arriving from $M!/(M-N)!N!$ if M equals fifty and N equals ten [56]. While this is certainly possible to accomplish and has the added benefit of guaranteeing that the optimal reduced data set is found, it is not a time efficient approach. This approach is not restricted to the spectral information of the data matrix; however, for simplicity the following discussion will focus only on the determination of the key set of spectra, because this is the method we used in our analysis. To determine the number of significant spectral factors (N) and to create a spectral initial guess for a curve resolution step,

IKSFA was applied to the two-way data matrix **X**. Keep in mind that in our work, the columns of **X** are a combination of the first and second dimension chromatograms and the sample injections (*M*) and the rows of **X** are the spectra (*L*) as described above; refer to Figure 3.5 for a schematic representation of the unfolded data and the data decomposition. IKSFA first decomposes the data using SVD equation (3.2). The search for a key set of orthogonal spectra uses the matrix **U** that contains the left singular vectors. Each row vector, $\mathbf{u_m}$, of the matrix, **U**, is first normalized to unit length

$$\tilde{\mathbf{u}}_{\mathrm{m}} = \frac{\mathbf{u}_m}{\left(\sum_{n=1}^{N} \mathbf{u}_{mn}^2\right)^{\frac{1}{2}}} \tag{3.3}$$

where $\tilde{\mathbf{u}}_{\mathrm{m}}$ is the normalized row vector, and the denominator is the norm of the row vector, since only the directions (row vectors that are perpendicular) and not the magnitudes of the row vectors are of interest in determining the most orthogonal rows. It is important to note that this is row-wise normalization as opposed to column-wise normalization.

The first key row corresponds to the row whose $\tilde{\mathbf{u}}_{m,1}$ value has the largest absolute value and we denote this row as $\tilde{\mathbf{u}}_{key1}$. This is a deviation from IKSFA as utilized by Schostack and Malinowski [56] where the first key row contained the minimum of the $\tilde{\mathbf{u}}_{m,1}$ value. This change was implemented due to the significance and uniqueness of the background spectra known to exist in the data set. A determinant is found for this key row and each remaining row, *r*, and the row with the maximum determinant

$$\max\left(\det\left|\begin{bmatrix} \tilde{\mathbf{u}}_{key1,1} & \tilde{\mathbf{u}}_{key1,2} \\ \tilde{\mathbf{u}}_{m,1} & \tilde{\mathbf{u}}_{m,2} \end{bmatrix}\right|\right) \tag{3.4}$$

is the second key row, $\tilde{\mathbf{u}}_{key2}$. This procedure is continued by adding a third row and finding the row that gives the maximum 3 x 3 determinant, etc., until *N* key rows are identified. It is at this point that iteration begins. The first key row, $\tilde{\mathbf{u}}_{key1}$ is replaced by the first row vector $\tilde{\mathbf{u}}_1$. If the absolute value of the determinant for the new key set is greater than that of the initial key set, the first row vector replaces the initial first key row; if the value is less than that of the initial key set, the key set remains unchanged, and $\tilde{\mathbf{u}}_{key1}$ is then replaced with the second row vector. This procedure is continued for the first key row for all *m* row vectors. The same logic is followed for all key rows completing one iteration cycle. Iteration continues until no change in the key set occurs after the completion of one complete iteration cycle [53, 57]. The key rows of **X**, *key1* through *keyN*, are then used as initial estimates for the componet spectra for the curve resolution algorithm, MCR-ALS, explained in the next section, as follows.

$$\mathbf{S}_{initialestimate} = \begin{bmatrix} \mathbf{x}_{key1} \\ \mathbf{x}_{key2} \\ \vdots \\ \mathbf{x}_{keyN} \end{bmatrix}^{\mathbf{T}} \tag{3.5}$$

3.4 Multivariate Curve Resolution - Alternating Least Squares (MCR-ALS)

MCR-ALS is a multivariate curve fitting technique that enables the analyst to mathematically separate chemical components in a data set by least squares optimization using knowledge of the data structure and the implementation of mathematical constraints that have a chemical significance [11]. These constraints can include nonnegativity, unimodality and multilinearity, among others. Equation 3.1 can be rearranged to solve for either the chromatographic matrix **R** (equation 3.6) or the spectral matrix **S** (equation 3.7) where the data matrix **X** is known. Either a spectral matrix estimate is used to solve for the chromatographic

matrix or a chromatographic initial estimate is used to solve for the spectral matrix. Orthogonal Projection Approach (OPA) [58], SIMPLE-to-use Self-modeling Mixture Analysis (SIMPLISMA) [59], and IKSFA (as used in this work to obtain a spectral initial estimate for MCR-ALS) are a few of the algorithms reported in the literature for the determination of either a spectral or chromatographic initial estimate. The MCR-ALS algorithm then iterates between equations 3.6 and 3.7 to minimize the error matrix until one of the two input iteration criteria is met; *i.e.*, until the fit error reaches a minimal improvement criterion or until a given maximum number of iterations has occurred [60]. If convergence to the global minimum occurs before the input iteration criteria are met, the least-squares model for the data set is found [61].

$$\mathbf{R = X\ (S^T S)^{-1}\ S} \tag{3.6}$$

$$\mathbf{S^T = (R^T R)^{-1}\ R^T\ X} \tag{3.7}$$

MCR-ALS decomposes the *n*-way data in a manner that produces *n* smaller matrices consisting of the pure component profiles for each dimension as shown in Figure 3.7 in which the chromatographic profiles are independently modeled [11]. This is important when the issue of tri- and quadrilinearity of the data matrix come into play. For example, PARAFAC, a factor analysis method, requires multilinearity due to the method used for the decomposition of the data and will be discussed further in section 3.5. In other words, to meet the requirements of multilinearity, the data matrix must be of the form where individual objects or samples (like variables) describe similar phenomena [9, 62]. Unfortunately, chromatographic data rarely meet this condition. Peak shifting and peak width variations are common causes for the lack of bi- or tri- linear HPLC data. Retention time drift is most often due to changes in column characteristics

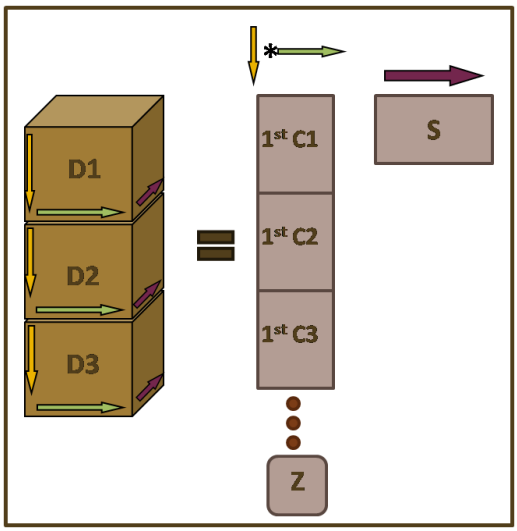

Figure 3.7: Visual representation of MCR-ALS data decomposition of three-way data to two-way data. Modified from reference [11]

(*i.e.*, stationary phase degradation) and uncontrolled minor changes in mobile phase composition during the chromatographic run, along with instrumental drift and interactions between analytes. [5, 6, 63] When chromatographic data are collected over long time periods, retention time shifts inevitably cause an increase in the complexity of the data due to misalignment of the detected peaks [63].

MCR-ALS has been shown to be applicable for *n*-way data where *n* can be two, three or four [60]. An advantage to our implementation of the MCR-ALS algorithm allows for the flexible implementation of chemically valid constraints for carrying out mathematical resolution of the data set reducing ambiguity in the model. These constraints, illustrated in Figure 3.8 are applied during the iteration procedure and work by eliminating mathematical solutions that are not chemically valid [60]. The non-negativity constraint prevents mathematically possible solutions that allow chemically invalid negative chromatographic, spectral or concentration responses. Unimodality constrains the resolved peaks by requiring that there be only one chromatographic maximum per compound and can be applied to every component in the mixture or a selected few. The spectral selectivity constraint allows for the restriction of portions of

selected spectral profiles. This constraint can be implemented to zero portions of spectral profiles that are known to not absorb above a given wavelength. The use of the trilinearity constraint implies that all of the chromatograms and the spectra of a pure component in a data set

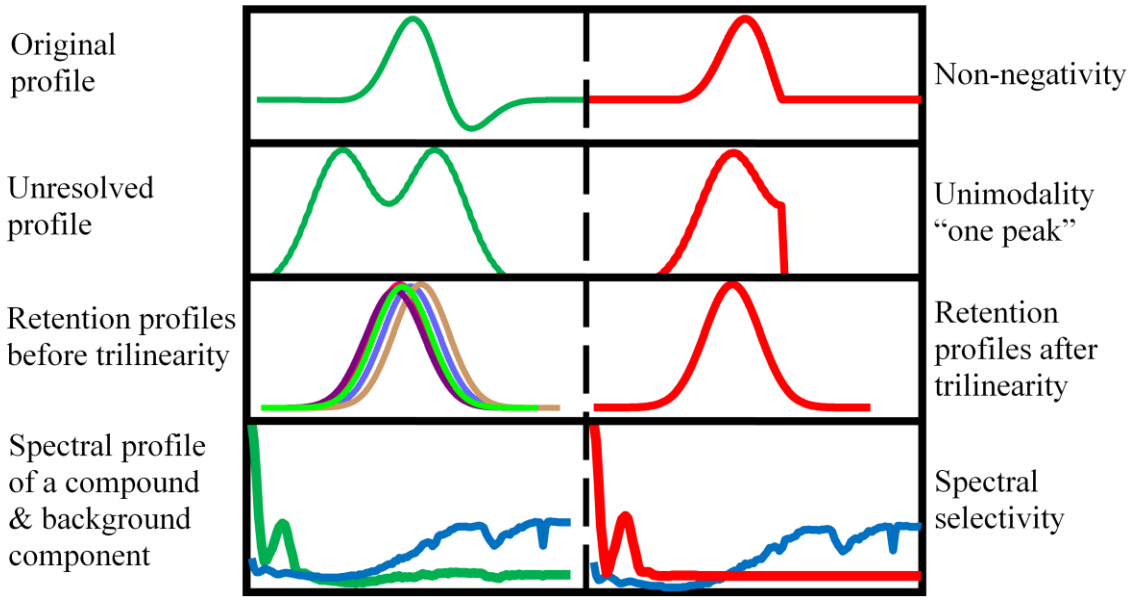

Figure 3.8: Constraints associated with MCR-ALS. From top to bottom: non-negativity, unimodality, trilinearity and spectral selectivity constraints. From left to right: before and after application of the corresponding constraint.

are identical. Many chemometric techniques for resolution of LC-DAD data require a trilinear structure. The MCR-ALS algorithm, however, does not require trilinearity in the data due to the method used to decompose matrix **X.** Instead, trilinearity is offered as an employable constraint if the data call for it. In general, the MCR-ALS technique follows the following steps:

1. Determination of the number of compounds present in the data matrix
2. Determination of an appropriate initial guess (chromatographic or spectral)
3. Determination of appropriate constraint inputs for the resolution procedure
4. Implementation of the optimized initial guess and constraint parameters until the iteration criteria are met [11].

The implementation of the IKSFA-ALS-ssel for a subsection of raw data results in the assignment of chromatographic peaks to their corresponding spectral components. An idealized representation of this is shown in Figure 3.9 for a four component model. The resolved **S** matrix has the dimensions of the number of wavelengths collected by four components ($L \times N$), while the resolved **R** matrix has the unfolded dimensions of the 1st and 2nd chromatographic dimensions and the number of samples by four components ($IJK \times N$). Samples 1 and 2 through K are shown, such that each spectral component of **S** corresponds to its color coordinated resolved chromatographic peak of **R**. The resolved **R** matrix is represented in two ways, first as a contour plot and then as the corresponding sequence of 2nd dimension chromatograms for each component. This illustration is idealized for simplicity and clarity, in that background component(s) are not represented, and each spectral component corresponds to a single, non-overlapped chromatographic peak. In the realm of real data, things are frequently not so straightforward. This point is clearly illustrated in section 5.4 by the results of the IKSFA-ALS-ssel analysis of a subsection of urine control data.

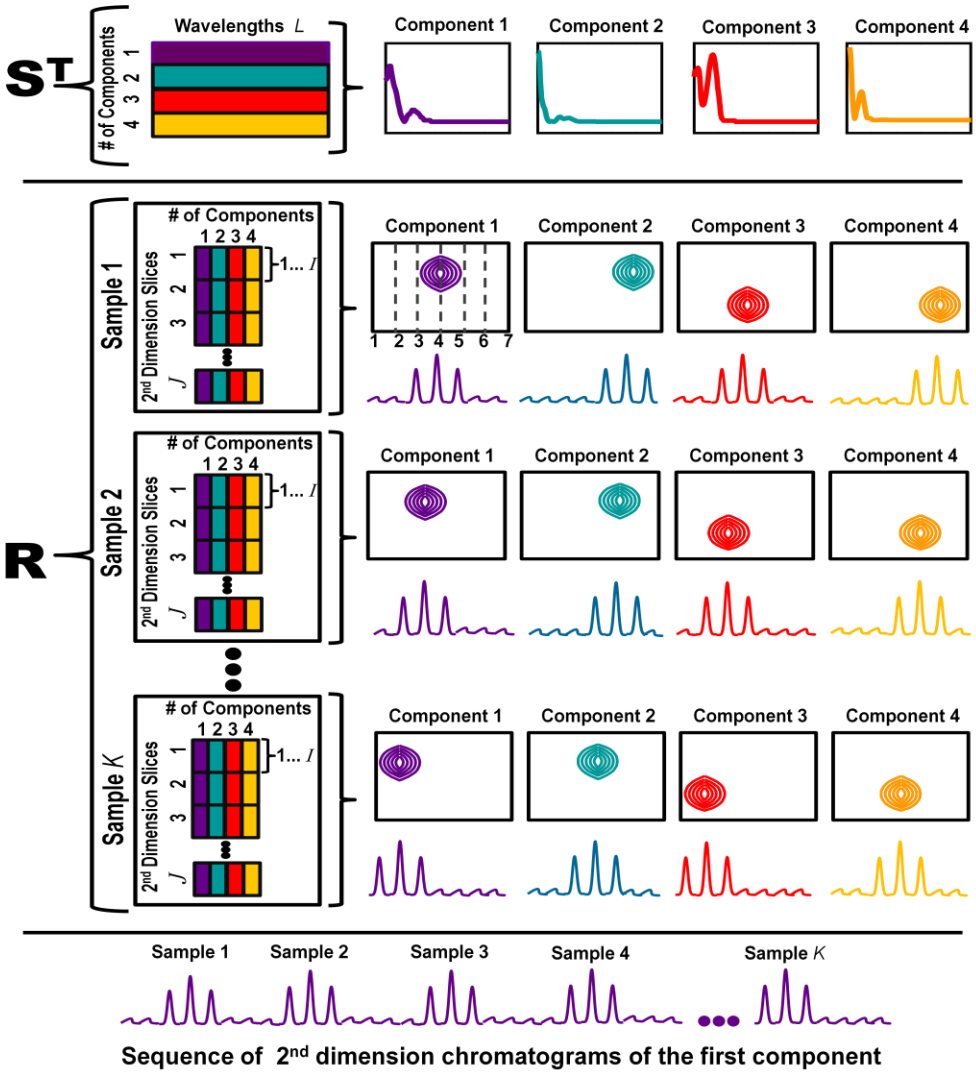

Figure 3.9: Schematic illustration of the resolution results from the IKSFA-ALS-ssel analysis in which both the data structure and corresponding graphical representations of the resolved spectra and chromatograms are shown. The data structure as shown consists of J=7 2nd dimension slices and K samples (1st dimension injections). Panel 1 illustrates the resolved spectral matrix. The dimensions of the resolved $\mathbf{S}^{\mathbf{T}}$ data matrix in this example are four spectral components by the total number of wavelengths collected (L). The corresponding spectra for each of the four components are plotted such that the y-axis is relative intensity versus wavelength. Panel 2 illustrates the resolved chromatographic matrix. Recall that each 1st chromatographic data point is equal to a 2nd dimension slice and that a 2nd dimension slice consists of I data points. The dimensions of the resolved **R** data matrix is unfolded to combine the 1st chromatographic dimension (J), the 2nd chromatographic dimension (I) and the number of samples (K) by the four spectral components. The four spectral components are color coordinated to correspond to their chromatographic counterpart. Each resolved chromatographic peak is graphically represented in two ways (1) by a contour plot of the 1st chromatographic dimension by the 2nd chromatographic dimension plotted for a given wavelength and a given sample and (2) a sequence of 2nd dimension chromatograms. Panel 3 shows the corresponding sequence of 2nd dimension chromatograms of component 1 for all samples (1−K).

3.5 Parallel Factor Analysis (PARAFAC)

The PARAFAC model, which defines decomposition of $\underline{\mathbf{X}}$ into multilinear components, was first introduced in the 1970s [64, 65], and can be expressed for four-way data sets such as are analyzed in our research as

$$x_{ijkl} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} d_{\ln} + e_{ijkl} \tag{3.8}$$

where a_{in} contains the second chromatographic dimension of component n in the i^{th} data point, b_{jn} is the first chromatographic dimension of component n at the j^{th} data point, c_{kn} is the relative concentration of component n at the k^{th} data point, d_{ln} contains the spectral information for component n at the l^{th} spectral data point and e_{ijkl} is the residual error term. These variables are the elements associated with the loading matrices **A**, **B**, **C** and **D** respectively. N is the total number of components in all of the samples [61, 66, 67]. The decomposition of the original data array $\underline{\mathbf{X}}$ can be accomplished using several different algorithms, such as direct trilinear decomposition (DTLD) [61], alternating trilinear decomposition (ATLD) [68], alternating slice-wise diagonalization (ASD) [69] and PARAFAC-alternating least squares (PARAFAC-ALS) [61]. The PARAFAC-ALS algorithm has the advantage of being capable of handling multi-way data, constrained models and missing data [61]. The disadvantage however, is the time required for the algorithm to perform hundreds or even thousands of iterations before the convergence criteria for uniqueness are met.

Fraga and Corley [42] reported for the first time the resolution and quantification of LC × LC data using GRAM followed by PARAFAC. They analyzed three different synthetic mixtures each containing a target analyte and at least one interferent: (A) *p*-chlorobenzoic acid and

benzoic acid (B) uracil and pyruvic acid (C) fumaric acid, maleic acid and phenyl phosphoric acid; where the target analytes are *p*-chlorobenzoic acid, uracil and fumaric acid. Both precision (as % RSD) and accuracy (as % bias) for each of the target compounds were reported and are listed in Table 3.1. The authors state that the lack of trilinearity due to retention time shifting significantly affects the performance of both the GRAM and PARAFAC algorithms even if retention time alignment is first performed on the data. This is a significant result driving our decision to use MCR-ALS as opposed to the PARAFAC model for analysis of the LC × LC data investigated.

Table 3.1 Quantitative results after resolution of the target analytes in each mixture studied by Fraga and Corley [42]

Compounds (target/interferent)	**Precision (% RSD)[a]**	**Accuracy (% bias)[b]**
p-Chlorobenzoic acid and benzoic acid	4.1	2.5
Uracil and pyruvic acid	21	2.8
Fumaric acid, maleic acid and phenyl phosphoric acid	12	66

[a] Percent relative standard deviation
[b] (Predicted concentration-true concentration)/true concentration

There are several other methods for the decomposition of multi-way data; Tucker3 [70] and unfolded PCA [71] are among the more common competitors of PARAFAC [10]. These methods are similar, in that, to achieve an accurate and condensed description of the original data they all decompose the data into scores and loadings. PARAFAC is, however, the simplest and the most restrictive of the algorithms. It can be thought of as a constrained version of Tucker3 in that it requires trilinear data to produce a unique, easily interpretable solution [10, 67, 71]. The obvious disadvantage here is that the multi-way data must first meet the criteria of being trilinear

or quadralinear (no retention time shifting can occur). However, the fact that the PARAFAC model will give a unique solution when trilinearity exists, means that the pure underlying spectra for a given set of components will be found since component rotation is not possible without a loss of fit [72, 73].

Chapter 4: Applicabiltiy of Chemometrics to Complex Samples

LC × LC has found applications in the fields of proteomics, pharmaceutics and metabolomics, where very complex samples are routinely analyzed, due to the increased peak capacity and therefore high resolving power of the technique. Metabolomics is the study of cellular processes by the chemical characterization of the low molecular weight compounds (metabolites) that are found in biological samples. Complex samples derived from genomics, proteomics and metabolomics studies are excellent candidates for analysis by fast LC × LC due to the demand in these fields for the resolution of a large number of constituents in such samples [3, 74]. Many of these biological samples are comprised not only of hundreds or thousands of constituents, but adding to the complexity of such samples, is the fact that those constituents have a concentration range that can exceed nine or ten orders of magnitude. As the concentration range of a given mixture increases, Nagels showed the necessity of also increasing the peak capacity to achieve the resolution of both the low and the high concentration constituents, thus making LC × LC ideally suited for such analysis [3]. A wide range of sample types (including cell cultures, microbes, plants and body fluids) have been used in metabolomic studies which involve the collection of quantiative data for the characterization of metabolites; *i.e.,* low molecular weight molecules [41, 75]. Our research to date has encompassed two of the above mentioned sample types for metabolomic studies: bodily fluids (urine) and plants (wine).

4.1 Urine and Standards Mixture Data

Many metabolomic studies involve complex biological fluids such as urine, blood and spinal fluid. The very nature of these fluids, unknown mixtures, implies limitations on the prior information available to the analyst at the time of data analysis, often making direct identification of chromatographically resolved peaks very difficult [54, 76]. Detection by means of MS/MS may allow for identification of those chromatographically resolved peaks. For this reason, many metabolomic studies involve only analysis of the identifiable major metabolites present in the sample. For example, LC and/or nuclear magnetic resonance (NMR) are techniques commonly employed in metabolomic profiling of urine centered around the identification of just a few known metabolites and/or pattern recognition studies [40]. The data from molecular profiling experiments allows for the screening of biomarkers to monitor the response of the body to drug treatment, surgery, or exposure to toxins by characterizing the changes in the small molecule metabolites present in urine [40, 74]. Since urine is especially sensitive to metabolic stressors such as disease or toxicity, it is a good sample for metabolic profiling [77].

Urine is replete in both endogenous and xenobiotic metabolites. The highly responsive nature of human urine to metabolic stressors such as disease or toxicity (a direct consequence of the body's autonomic response to eliminate substances in an attempt to maintain homeostasis) offers several overwhelming advantages [77, 78]. Due to this autonomic response, detection of the changes in the concentrations of the endogenous metabolites in urine has the potential to increase our understanding of the mechanisms of disease and drug action; and detection of the changes in the xenobiotic metabolites in urine has the potential to aid in the discovery of biomarkers for drug efficacy and toxicity and of biomarkers for disease risk [41, 79, 80]. LC

and/or NMR are techniques commonly employed in metabolomic profiling of urine centered around the identification of just a few known metabolites and/or the use of pattern recognition techniques applied to unidentified signals [40]. Non-targeted, global profiling of metabolites in human urine has been accomplished in recent studies using gas chromatography mass spectrometry (GS-MS) [79, 81, 82].

A comprehensive LC × LC system was developed by Stoll and Carr at the University of Minnesota. The chromatographic separation of the urine samples analyzed in this research was performed in the laboratory of Dr. Carr [22, 83]. The system employs the use of a dual gradient and the use of high temperature in the second dimension. This was accomplished through the use of an eluent preheater and a heating jacket placed around the second dimension reversed-phased carbon-clad zirconia column. The sample to be injected onto the first dimension system is preheated to 40 ^{o}C before passing through the first column. The chromatographic conditions for the first dimension are as follows: gradient elution from 0 to 70% B from 0 to 23 min, where A is 20 mM sodium phosphate, 0.1 mM EDTA at pH 6 and B is acetonitrile, with a flow rate of 0.1 mL/min. The ^{1}D stationary phase is a lab-made hydroxylated-hypercrosslinked material [15, 84] packed in a 200 mm x 1.0 mm column. This column has a benzylic hydroxyl functionality embedded into the hyper-crosslinked platform, allowing the relatively polar stationary phase to be employed in a reverse phase manner requiring a much weaker mobile phase than needed for a C-18 column. This provides compatibility with the second dimension mobile phases to reduce peak broadening [84]. Effluent from the first column is captured by a 10-port valve in 21-second fractions, a 35 µL sample. Two pumps are used to sequentially deliver the aliquots loaded in the two 35 µL loops in an alternating fashion to the second dimension column, where the ^{2}D stationary phase is a carbon-clad zirconia material packed in a 33 mm x 2.1 mm The second

dimension column is maintained at 110 ^{o}C. This allows the use of high flow rates, effectively reducing the time required for the second dimension separation. The chromatographic conditions for the second dimension column are as follows: gradient elution from 0 to 100% from 0 to 17.45 s, where A is 20 mM phosphoric acid and B is acetonitrile, with a flow rate of 3 mL/min. The re-equilibration time was 3 seconds. Diode array detection (DAD) from 200 nm to 700 nm was employed after the second separation column and the data was recorded every four nm [22].

Fourteen injections of urine control sample (Figure 4.1A) and six injections of a standards mixture (Figure 4.1B) were interspersed over the course of a 64 injection experiment using the above described LC × LC system [22, 83] requiring over thirty hours of total run time. The standards mixture consists of nitrate, tryptophan, hydroxytryptophan, indole-3-acetic acid, indole-3-propionic acid, indole-3-acetonitrile and tyrosine. The acquired data consist of absorbance values in mAU units as the dependent variable, and the independent variables being retention on the first dimension column, retention on the second dimension column, UV-visible wavelength, and sample injection number. Thus, for the standards mixture replicates, the size of the array is 840 × 84 × 6 ×126 (2nd chromatographic dimension, 1st chromatographic dimension, number of sample injections, wavelength from 200 nm to 700 nm at 4 nm intervals) and the size of the array for the analyzed section for urine control data is 161 × 26 ×14 × 126. The data were imported into the MATLAB environment using ACDLABS ChromProcessor 9.0 (Advanced Chemistry Development, Inc. Toronto, Canada). The data were analyzed using MATLAB software R2007a (Mathworks, Inc. Natick, MA) and a HP Pavilion dv9500 with 4GB RAM, an Intel® Core™ 2 Duo CPU T7500 @2.20GHz processor operating with the Windows Vista Home Premium operating system. The MCR-ALS algorithm used for this analysis has been

described previously by this group [60]. LCImage software (GC Image, LLC Lincoln, NE) was provided by S. Reichenbach [22]. The results of the analysis of these data are described in Chapters 5 and 7.

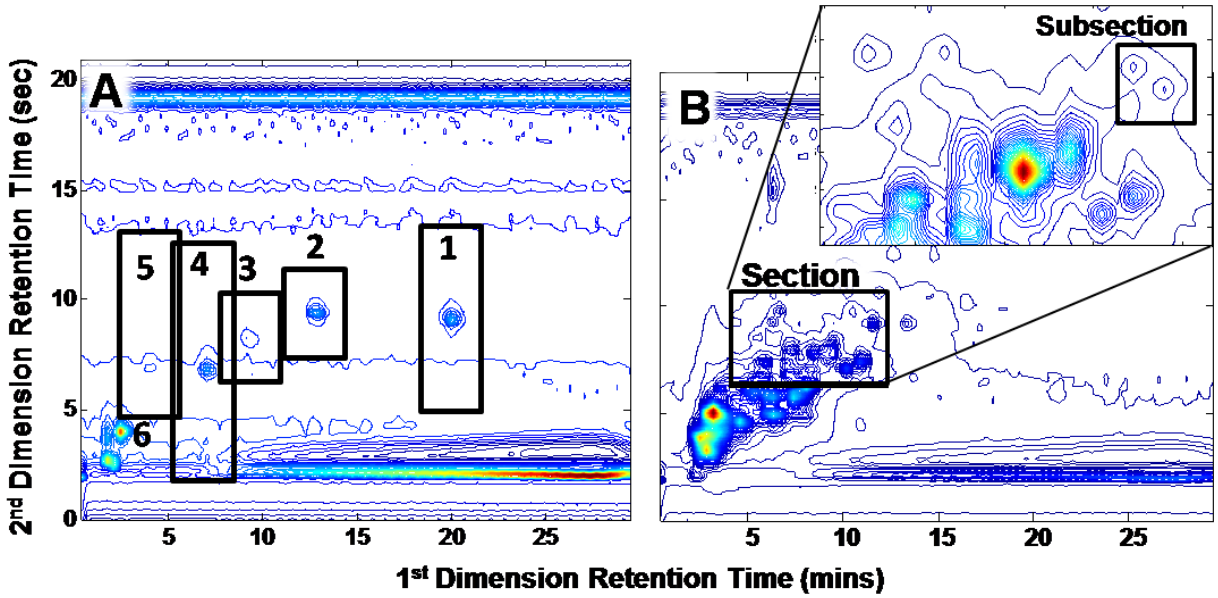

Figure 4.1: Contour plots at 216 nm of two different sample types within the 64 injection 2D-LC-DAD run. (A) Contour plot of the third replicate injection of the standards mixture where peak 1 is indole-3-acetonitrile, peak 2 is indole-3-propionic acid, peak 3 is indole-3-acetic acid, peak 4 is tryptophan peak 5 is hydroxytryptophan and peak 6 is tyrosine. (B) Contour plot of the seventh replicate injection of the urine control standard. Inset shows the section of data selected for chemometric analysis.

4.2 Wine

Wine consists of several thousand compounds of varying concentrations [85, 86]. The major components of wine are water, ethanol, glycerol, sugars, organic acids and various ions. The minor components include amino acids, aliphatic and aromatic alcohols, and phenolic compounds such as anthocyanins, flavonols and catechins [85, 87]. Analytical analysis of wine is frequently performed for quality control, compound identification and authenticity studies [88, 89]. Since it is one of the most ingested beverages in the world, quality control of the product is critical [90]. Authenticity studies are also quite critical in the determination of vineyard

geography (origin of the wine), vine variety (types of grapes used in the production of the wine) and age (length of fermentation of the wine) [85].

Due to the complex nature of wine, classification, screening and compound identification can be tedious and time consuming. Many classification methods, such as principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) have been used in the differentiation of wines based on geographical origin and grape variety [89, 91, 92]. Most often the discrimination is based on a limited number of targeted compounds. In the case of wines, Capron *et al.* used 63 parameters (concentrations of trace elements, macro elements, biogenic amines, glycerol and malic acid along with the ratios of isotopes) to classify wines from four different countries (South Africa, Hungary, Romania and Czech Republic) [92]. The Wine Database European Project analyzed, resolved and quantified the concentration and ratio values for the 63 parameters associated with the 393 analyzed wines [92]. While the models studied in this work were all successful at discrimination of the four geographic locations, the authors make the following important statement. "It must be underlined that the models described in this article are built for the first vintage year of the project. Since wines are depending on the vintage, it is probable that models presented here must be updated in order to deliver the same quality prediction." This implies an extensive amount of work to be accomplished on a yearly basis. Markris *et al.* analyzed nineteen polyphenolic compounds for each of forty wine samples using HPLC-DAD (140 minute run time) followed by discriminate analysis (DA). Discrimination of both geographical location and cultivar was found to be possible. While geographical discrimination required seventeen of the nineteen quantified polyphenols, only eleven were required for the discrimination of cultivars. In other words, eight peaks in forty samples (320 peaks) were quantified unnecessarily for this type of targeted discrimination.

The wine samples for this research were acquired from three different vineyards: University of Minnesota Horticultural Research Center (HRC), Winter Vineyard (WV) and Ceplecha Vineyard (CV) [3]. The experimental work for these samples was carried out in the lab of Professor Peter W. Carr at the University of Minnesota [3]. Three different HRC samples were acquired from the same five day fermentation batch to act as control samples or to investigate within batch variability. Both the WV and CV samples were acquired from a five day fermentation batch to investigate geographical variability. Three replicate injections were analyzed for all of the above samples. All samples were fractionated by size-exclusion chromatography in order to remove the carbohydrates and other large molecules. A small molecule fraction was collected and evaporated to dryness to remove all of the ethanol. Samples were reconstituted using the first dimension mobile phase and 40 µL were injected onto the first dimension column of the comprehensive two dimensional liquid chromatographic (LC × LC) system developed by Stoll and Carr at the University of Minnesota [3]. The chromatographic conditions for the first dimension column are as follows: flow rate of 0.10 mL/min, gradient elution for 0 to 50 % B from 0 to 23 minutes. Mobile phase A consists of 20mM sodium dihydrogen phosphate, 20 mM sodium perchlorate and 0.2mM EDTA at a pH =5.7. Mobile phase B is acetonitrile. The column was a Discovery HS-F5 100mm x 2.1 mm. Small aliquots were collected in two loops and the contents of each loop were sequentially injected onto the second dimension column which was maintained at 110 ^{o}C. The second dimension cycle time was 21 seconds. The chromatographic conditions for the second dimension column are as follows: gradient elution from 0 to 100% from 0 to 17.45 s, where A is 20 mM phosphoric acid and B is acetonitrile, with a flow rate of 3 mL/min. The re-equilibration time was 3 seconds. Diode array detection was employed from 200 to 700 nm after the second dimension column.

The acquired four-way data consists of a second retention time dimension, a first retention time dimension, sample injections, wavelength (independent variables) and absorbance values in mAU units as the dependent variable, 840×84×15×126 for the geographical comparison. Figure 4.2 is a representative contour plot of a wine sample and the box indicates the section of data that was analyzed. The results of the wine analysis are described in Chapter 8.

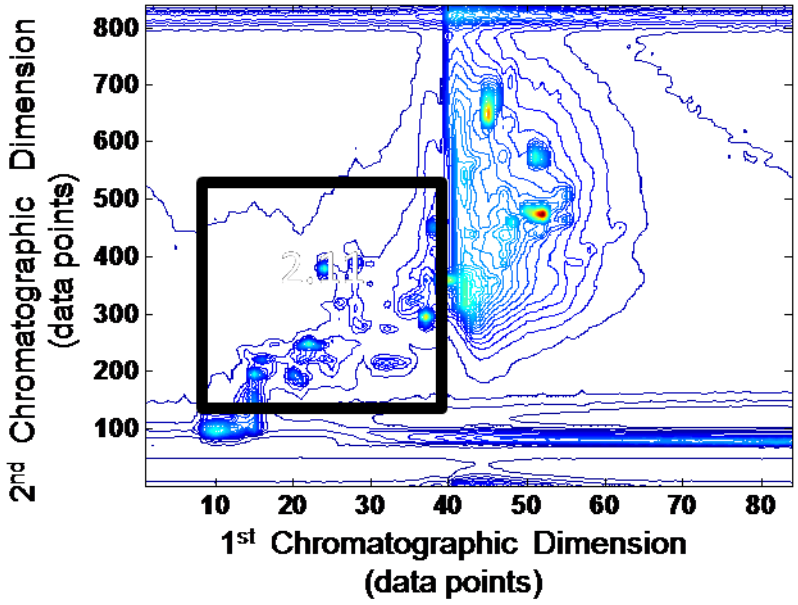

Figure 4.2: Contour plot of HRC C 1st replicate at 216 nm. The boxed area is the section of the chromatograms analyzed in this work.

4.3 Phenytoin in wastewater samples

The anticonvulsant phenytoin is a widely prescribed first-line anti-epileptic (AED) drug. Phenytoin is known to have serious effects on bone mineral density, to cause AED-related cutaneous adverse reactions, and to cause birth defects; *i.e.*, it is a teratogen [93-95]. Pharmaceuticals and personal care products (PPCPs), such as phenytoin, are emerging as an important class of pollutants that can be found in surface and ground water and in sewage effluents acquired from wastewater treatment plants [96-98]. This increased attention is due to several different concerns regarding their effects on both human and wildlife, such as bacterial

resistance to antibiotics and estrogenic effects [99]. The ability to accurately and precisely detect and quantify these types of contaminants is of the utmost importance in the determination of the potential human health risks and environmental risks. The current most popular method for analytical analysis of PPCPs is liquid chromatography with tandem mass spectrometry (LC/MS/MS); however, ultra high performance liquid chromatography-time-of-flight-MS (UHPLC-TOF-MS) is becoming more widely utilized [99]. One disadvantage associated with these techniques is the high cost of the instrumentation itself [100].

The experimental work for these samples was carried out in the laboratory of Prof. Stoll at Gustavus Adolphus College [27, 101]. Solid Phase Extraction (SPE) was used to pre-concentrate 16 L of urban wastewater treatment plant effluent (WWTPE) samples to a 16 mL sample yielding a pre-concentration factor of 1000-fold. This sample was used to prepare 10 sample injections, of WWTPE, each spiked with phenytoin as follows: no spike, 25, 50, 75, 150 parts-per-billion (equivalent concentrations of 25, 50, 75, and 150 parts-per-trillion in the original samples prior to extraction); each sample was injected twice. A series of phenytoin standards at the same concentrations was also prepared in distilled water (DI). The developed three-dimensional separation was utilized such that the sLC × C method followed a LC heart-cut ^{1}D run, due to the complexity of the samples. To illustrate said sample complexity of the WWTPE samples, Figure 4.3 [102] shows the targeted three-dimensional chromatographic analysis of a similar WWTPE sample after each of three chromatographic separations. This work was performed in the Stoll group prior to the sLC × LC analysis of WWTPE samples that are the topic of this work. As is clear from Figure 4.3 A, resolution of the sample constituents has clearly not occurred after the 1D-LC separation. By hearting-cutting the first dimension effluent, an improvement in resolution is seen in Figure 4.3 B; however, quantification of the

targeted peaks is still not achievable for the target compounds of phenytoin and chlorphene. A second heart-cut procedure is performed and sent to the third column where a resolution is achieved and quantification is now possible, as shown in Figure 4.3 C.

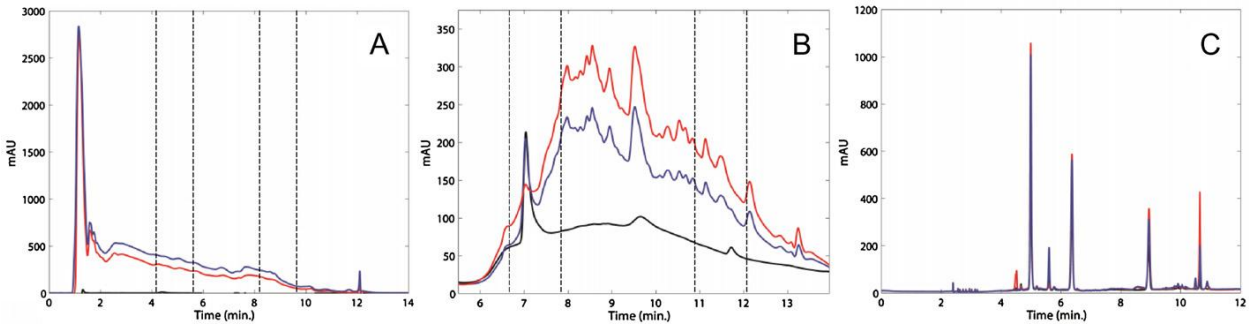

Figure 4.3: Chromatograms observed at the outlet of each dimension of separation in the 3DLC/UV system for the separation of the 1000-fold concentrated WWTP effluent sample. The red chromatogram: neat WWTPE extract, blue chromatogram: phenytoin and chlorophene standards spiked into the WWTPE at 500 and 50 ppb, black chromatogram: phenytoin and chlorophene standards spiked in DI water at 500 and 50 ppb. Reproduced from reference [101] with permission from Elsevier.

Specific to this work, a heartcut portion of the effluent (between 2.3 and 3.5 minutes) was transferred from the ^{1}D column, which was an Ascentis Express F5 perfluorophenyl stationary phase (75 mm x 2.1 mm i.d.) to the ^{2}D column (a serially-coupled pair of 50 mm x 2.1 mm i.d. column prepared in-house with carbon-modified silica: 15 % carbon w/w, United Science, LLC, Minneapolis, MN). The ^{2}D separation was isocratic using 40/60 ACN/10 mF H_3PO_4 with a flow rate of 0.5 mL/min. Six 2-second fractions of ^{2}D column effluent (between 7.82 and 8.02 minutes) were stored in six valve loops for consecutive injection onto the ^{3}D column (Ascentis Express C18, 30 mm x 2.1 mm i.d.). The ^{3}D analysis was a 20-second isocratic separation for each of the six ^{2}D separation fractions where the eluent was 25/75 ACN/10 mF H_3PO_4, with a flow rate of 2.0 mL/min and maintained at 50^{o}C. Both the ^{1}D and ^{2}D effluents were diluted with DI water, for pre-column focusing, at a flow rate of 0.5 mL/min and the columns were maintained at 40^{o}C. Absorption of UV and visible light was detected using a DAD in the range

of 200 to 800 nm, at 4 nm increments. Prior to chemometric analysis, the data set was sectioned to encompass only the region containing the phenytoin and interferent peaks, as shown in Figure 4.4 where the shaded portions of the contour plots were eliminated from the data analysis. The results of the phenytoin analysis are described in Chapter 6.

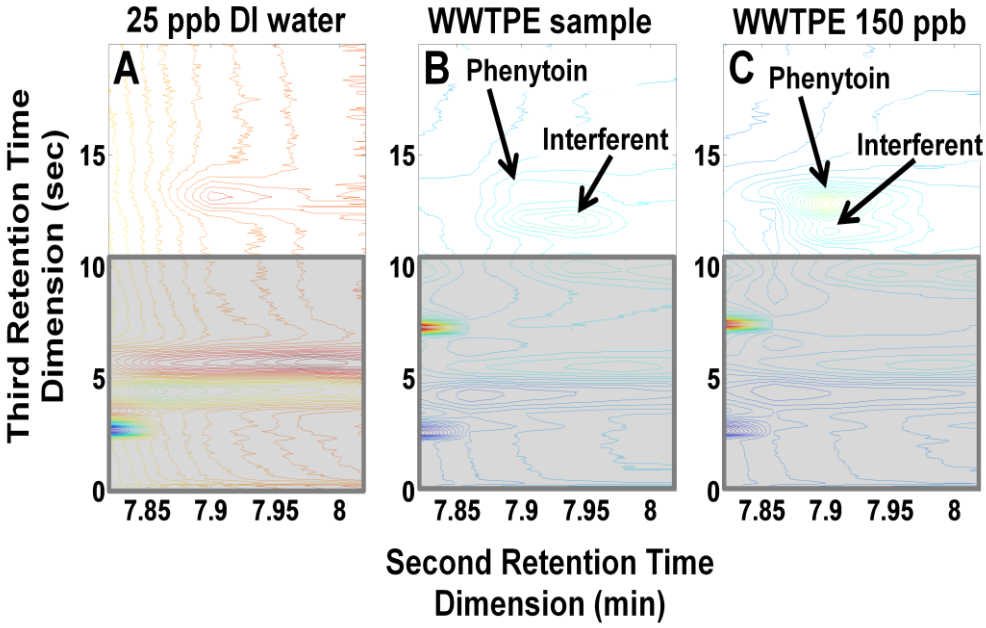

Figure 4.4: Contour plots of various sample injections at 216 nm for the phenytoin study before chemometric analysis. The shaded portion of the plots is the section of the data eliminated from the chemometric analysis of the data. (A) Contour plot of DI water sample spiked with 25 ppb phenytoin. (B) Contour plot of the WWTPE sample without a spiked amount of phenytoin. (C) Contour plot of the WWTPE sample spiked with 150 ppb phenytoin.

Chapter 5: Chemometric Resolution and Quantification of Four-Way Data Arising from Comprehensive 2D-LC-DAD Analysis of Human Urine

Adapted from H.P. Bailey, S.C. Rutan, Chemom. Intell. Lab. Sys., 106 (2011) 131-141.

The need for chemometric methods capable of resolving and quantifying data arising from LC × LC separations of complex samples is ever more urgent in order to obtain the maximum information available from the data. To this end, a chemometric method was developed that combines iterative key set factor analysis and multivariate curve resolution-alternating least squares analysis with a spectral selectivity constraint. The work in this chapter details the analysis scheme, explores both the standards mixture and urine control data and shows this method to be capable of resolving chromatographically rank deficient, non-multilinear data. (Spectrally rank deficient compounds can only be quantified if the peaks having the same spectra are chromatographically resolved.) Over 50 chromatographic peaks were found in a relatively small section of a LC × LC-diode array data set of replicate urine samples (a four-way data set) using the developed method. The relative concentrations for 34 of the 50 peaks were determined with % RSD values ranging from 0.09 % to 16 %.

5.1 Quantification Algorithm Development (relative concentration determination)

In LC × LC, a first dimension peak consists of several second dimension injections (slices across a first dimension peak consisting of *J* data points). Each second dimension

injection (slice) will produce a second dimension chromatogram consisting of *I* data points. This data structure is specifically discussed in section 3.1 and illustrated in Figure 3.9. The quantification algorithm is based on the premise that the sum of these second dimension peak areas is equivalent to the volume of that LC × LC peak [48, 50]. Figure 5.1 A illustrates this premise, in which the same single compound is present in six replicate sample injections. It can be seen that sample injection 1 consists of four sequential second dimension peaks. The areas under each of these four second dimension peaks are determined and are summed in order to ascertain the LC × LC peak intensity [83]. This procedure is followed for all six sample injections shown in Figure 5.1A and allows for the comparison of the relative concentrations of the single compound present in all six sample injections. This method will be referred to as the manual baseline method throughout this work and is equivalent to the area summation method described by Thekkudan *et al*. [50].

For simplicity, Figure 5.1A represents an ideal case in which the peak has been well resolved from the background components using the developed chemometric method and only one compound is present in the section of the data analyzed, as opposed to Figure 5.1B, which shows a plot of the corresponding raw data. Unfortunately, the ideal case is not frequently observed and it is extremely likely that other components of the sample may have the same first dimension retention time but an earlier or later second dimension retention time. In such an instance, additional second dimension peaks will elute in the individual second dimension chromatograms either before or after the peak of interest. Any second dimension peaks not associated with the peak of interest are simply left unintegrated and thus do not contribute to the relative concentration calculation. For very simple mixtures, it is straightforward to determine

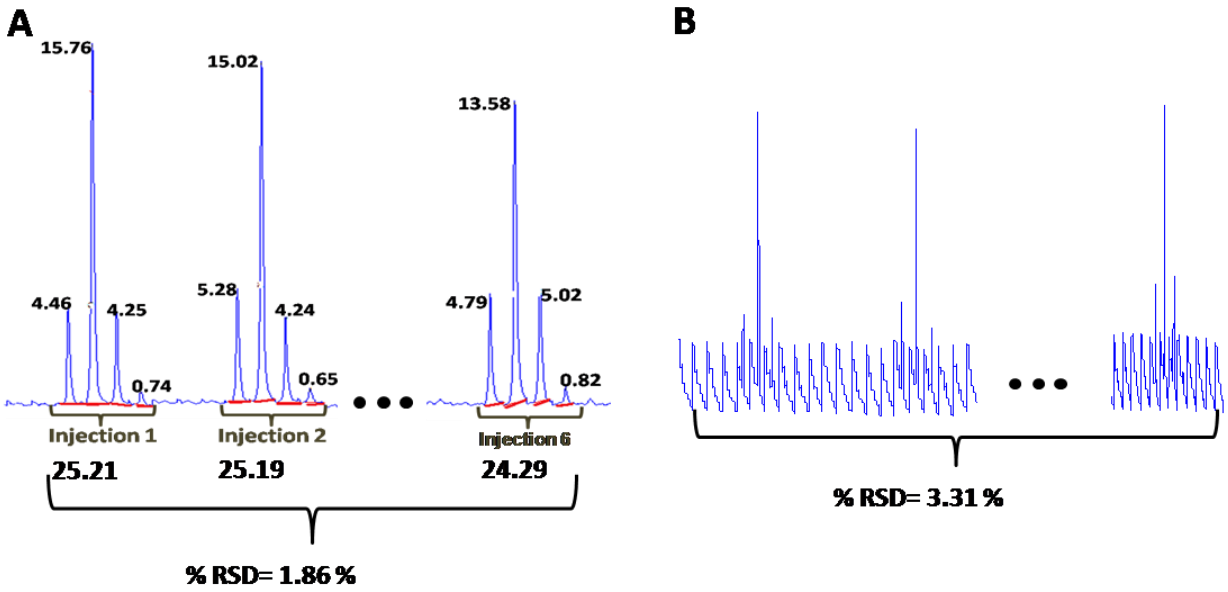

Figure 5.1: Comparison of the subsection for the standards mixture containing Peak 1 showing the sequence of resolved 2nd dimension chromatograms and the raw data for the injection of six replicate samples onto the 1st dimension column. (A) The data after application of the developed chemometric method. The line drawn under each 2nd dimension peak shows the manually determined baseline, and the areas for each of the second dimension peaks (shown at the top of the peaks) are totaled (shown at the bottom of each peak grouping), giving the relative concentrations of Peak 1 for each of the six sample injections and the % RSD showing the precision of the quantification. (B) A plot of the sequenced second dimension chromatograms of the raw data. Each 1st dimension sample injection gives rise to seven 2nd dimension injections with four of those injections containing the peak of interest and three injections consisting only of the background in this example.

which second dimension peaks comprise a given LC × LC peak, but for more complex mixtures, that are of interest in the present work, this can be challenging at best. The IKSFA/MCR-ALS curve resolution procedure is used in the present work to resolve all spectrally distinct components into individual LC × LC chromatograms, which are much simpler, and can therefore be more easily and precisely integrated using the above procedure. Often the method resolves weakly absorbing peaks that were not visually observable in the raw data and may not be spectrally distinct. The advantage of the manual baseline method is that spectral uniqueness is not necessary as long as the peaks in question are chromatographically resolved and a manual baseline can be drawn for integration.

5.2 Data Analysis Scheme

The LC × LC-DAD data used in this work did not possess the required multilinearity condition (*i.e.*, no retention time shifting, highly reproducible chromatographic peak shapes and consistent spectral responses) needed to implement certain chemometric techniques such as PARAFAC and GRAM; thus we opted to employ multivariate curve resolution (MCR) techniques in the analysis of the LC × LC-DAD data. It is important to note, there is not, in either retention time dimension, a common aligning factor to which all peaks can be shifted. A section of the data where the absorbance was less than two was chosen for chemometric analysis. This was done in order to assure a linear relationship between absorbance and concentration as stated by Beer's law. Due to the complexity and size of the chosen data section, the data were further divided into subsections. The data analysis procedure followed for the analysis of both the standards mixture data and the urine control data is outlined in Figure 5.2. Subsections were initially determined by creating contour plots to determine the 1^{st} and 2^{nd} dimension data point boundaries around a visually observable peak. Once a subsection was created, chemometric data analysis began with SVD and IKSFA of the data matrix **X** (dimensions $IJK \times L$). The initial input parameter for IKSFA, (in this work the number of spectral components (N) as opposed to the number of chromatographic components), was to some extent subjectively determined using a combination of two visualization methods, a scree plot and a contour plot of the subsection to be analyzed. The initial N spectral components obtained from the IKSFA analysis are then used as a spectral initial estimate ($\mathbf{S}^{\mathbf{T}}$) for the initialization of the in-house MCR-ALS algorithm [60] which employs the non-negativity constraint in the chromatographic dimension. These two steps are repeated for several different possible numbers of components to ensure that as many

possible components are found without over-fitting the data (see section 5.6 for a more detailed description).

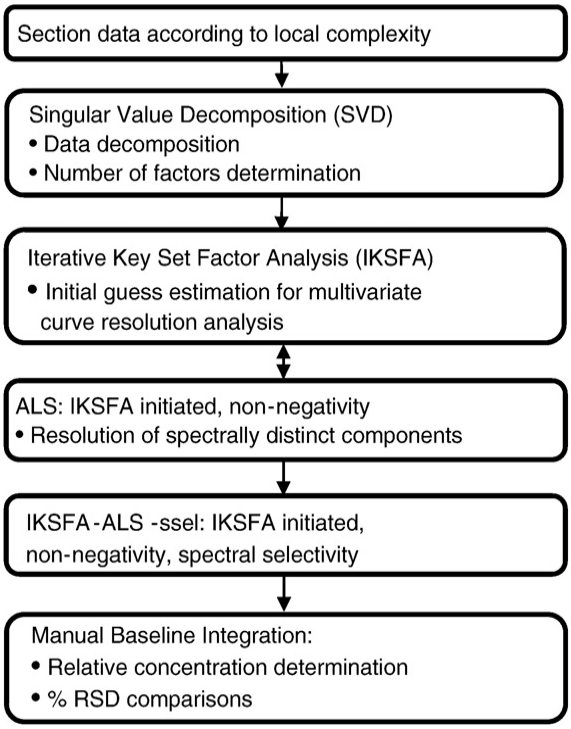

Figure 5.2: Chemometric data analysis scheme used in the resolution and quantification of LC × LC data.

After the optimization of the number of components, a final MCR-ALS analysis step, referred to from here on as ALS-ssel, where ssel denotes the use of the spectral selectivity constraint, is performed in which three constraints are applied to the analysis:
(1) chromatographic non-negativity, applied as in the previous analysis steps. (2) spectral selectivity, and (3) spectral non-negativity. Our implementation of spectral selectivity, constrains only the non-background components so that the last 51 spectral data points (corresponding to wavelengths 440 nm to 700 nm) were set to zero. The spectral non-negativity constraint was selectively applied to correspond to the parameters of the spectral selectivity constraint so that the background components are allowed to be negative but the compound

spectra are constrained to be greater than or equal to zero. The components that require the application of constraints (2) and (3) were identified in the first MCR step.

The implementation of the IKSFA-ALS-ssel approach described above for a subsection of raw data results in the assignment of chromatographic peaks to their corresponding spectral components. The spectral component that contains the peak of interest is further analyzed to determine the relative concentrations of the resolved peak for each sample injection, and the % RSD was then calculated. This is accomplished by plotting the resolved chromatographic results for only the component of interest and for a given sample injection as a sequence of second dimension chromatograms. This allows for good baseline visualization of the compound of interest in a given injection for implementation of the manual baseline method as was previously described in section 5.1. After the manual baseline method has been utilized to determine the relative concentrations of the compound of interest, the % RSD for that peak was determined by dividing the standard deviation of the replicate sample injections by the average determined relative concentrations for all replicate sample injections and multiplying by 100. Due to the data structure (replicate injections without calibration injections) it was not possible to calculate the accuracy of the method; only the precision of the method can be discussed.

The above described procedure was followed for the eighteen data subsections that were created for the eighteen visually observable peaks. The analysis of these eighteen subsections revealed additional peaks not previously observed in the raw data contour plot of the entire section. New subsections were created for the analysis of the previously unobserved peaks as these peaks were detected, so that both observed and initially undetected peaks in the data were appropriately analyzed. A full discussion of the choices and reasoning behind why the above steps were undertaken is found in section 5.6.

5.3 Standards Mixture Analysis (effects of subsection size and number of components)

The six replicate standard mixture injections (Figure 5.3A and also described in detail in Chapter 4.1) were interspersed throughout a 64 injection run and contained six known compounds that were intended to be well resolved. However, multiple contaminants were found in close proximity to Peak 6 for all of the replicate sample injections. Therefore, this peak was not included in the following analyses, because the goal of this portion of the work was to limit possible interfering variables, such as chromatographic and spectral rank deficiencies, to obtain a better understanding of how the algorithm functions. The % RSD values for the concentrations of Peaks 1-5 for the raw data using the manual baseline method 9as described in section 5.1) and the chemometrically resolved data using both the manual baseline method and LCImage software (refer to section 2.6) volume determination are shown in Table 5.1. For clarity, it is

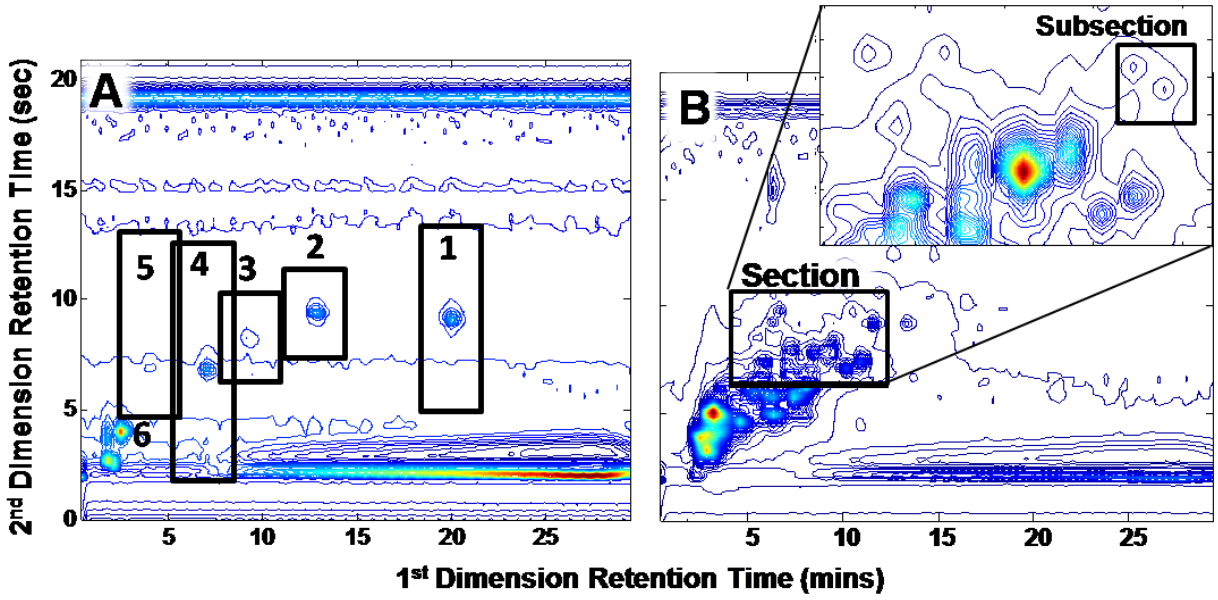

Figure 5.3: Contour plots at 216 nm of two different sample types within the 64 injection 2D-LC-DAD run. (A) Contour plot of the third replicate injection of the standards mixture where peak 1 is indole-3-acetonitrile, peak 2 is indole-3-propionic acid, peak 3 is indole-3-acetic acid, peak 4 is tryptophan peak 5 is hydroxytryptophan and peak 6 is tyrosine. (B) Contour plot of the seventh replicate injection of the urine control standard. Inset shows the section of data selected for chemometric analysis.

important to note that the subsection sizes used for the analysis of both the raw data and chemometrically analyzed data are those found to be the optimal subsection sizes for each peak to be further discussed below. Also, the LCImage software does not require the use of subsections around individual peaks of interest. Peaks 3 and 4 (indole-3-acetic acid and tryptophan as shown in Figure 5.3A) give the highest % RSD values for both data types and quantification methods. Peak 3 is a very weakly absorbing compound making it difficult to accurately determine the peak baseline from the high background in the raw data. The reason for the high % RSD for Peak 4 is that there is an overlapping contaminant found to be present only in injection 2. The IKSFA-ALS-ssel resolved data yields better results as compared with the raw, unresolved data, except for Peak 2. Overall, there is an average three-fold improvement in precision over integration of the raw data.

Table 5.1: % RSD results for the precision of peak quantification of both raw and IKSFA-ALS-ssel resolved data of the standards mixture injections for Peak 1 through Peak 5.

Standards Mixture Data	**Manual Baseline**		**LC Image software**
	Raw Data	IKSFA-ALS-ssel	IKSFA-ALS-ssel
Peak 1	3.31	1.60	9.07
Peak 2	1.75	2.16	10.5
Peak 3	12.6	4.71	34.5
Peak 4	13.1	3.47	19.4
Peak 5	5.21	1.40	1.30
Ave % RSD	6.53	2.61	15.0

The analysis of Peaks 1-5 using IKSFA-ALS-ssel for different subsection sizes was done to determine whether the size of the subsection chosen to encompass the peak of interest would have an effect on quantification. Due to large retention time shifting in the first retention time dimension, it is important that the subsection include all data points which reflect the presence of

the compound of interest; however, after this criteria is met, is it in the best interest of the analysis for the subsection size to be small (just encompassing the peak of interest), as large as possible (allowing for additional data points that might allow for more accurate determination of the background component) or does subsection size have any effect on the % RSD values at all? Table 5.2 gives the % RSD values as determined after IKSFA-ALS-ssel analysis using the manual baseline method for five different subsection sizes for each of the Peaks 1-5. The first and second dimension coordinates for the maxima of the Peaks 1-5 were visually determined, and the peak was centered within each subsection so that the first dimension for all subsection sizes contained ten data points. This range of points in the first dimension ensured that the peaks are not cut off in the first elution time dimension. The number of data points in the second dimension for the five different subsection sizes were 200, 250, 300, 350 and 400 data points respectively (200 data points was chosen as a minimum subsection size because smaller subsections led to the peak being cut off in peaks 2-4). From Table 5.2 it is clear that as the subsection size increases, the % RSD decreases until a critical limit is reached, at which point the % RSD increases with increasing subsection size. This trend is directly related to the signal to noise ratio of the given subsection size for a specific component. We conclude that for smaller subsections, there are two contributing issues that lead to the higher % RSD values. For one, if the peak is large relative to the background component (such that the peak "overwhelms" the size of the subsection) the analysis method will have difficulty in accurately estimating the background contribution. Second, upon integration, the lack of data points on either side of the peak in the resolved sequenced chromatogram makes a consistent baseline determination more difficult. For the larger subsections, the issue is the opposite, particularly for weaker peaks; *i.e.*, the method has difficulty in accurately estimating the peak contribution. In other words, the peak

Table 5.2: Effects of subsection size on the % RSDs of IKSFA-ALS-ssel analyzed standard mixture data for Peak 1–Peak 5.

Size of subsection in data points	% RSD Values				
	Peak 1	Peak 2	Peak 3	Peak 4	Peak 5
Subsection 1 (200 x 10)	2.71	**2.16**	**4.71**	CO	5.04
Subsection 2 (250 x 10)	2.22	2.30	7.76	5.33	4.08
Subsection 2 (300 x 10)	2.03	2.31	12.67	5.85	2.88
Subsection 3 (350 x 10)	**1.60**	2.68	NA	4.93	**1.40**
Subsection 4 (400 x 10)	1.86	7.59	NA	**3.47**	2.69

CO: the peak was clearly cut off for this subsection size.
NA: no available results due to very low peak to background ratio
Entries in bold denote the optimal subsection size.

gets lost in the background. This is especially evident in Peak 3 for subsection sizes 4 and 5 in which the background was so large in comparison to the weak peak that the algorithm was unable to yield a resolution of the peak that was quantifiable. Therefore, the most appropriate subsection size is dependent on the relative intensity of the target compound within the subsection.

5.4 Urine Control Sample Analysis (curve resolution and quantification)

Over fifty peaks were found within the section of the urine control chromatogram that was analyzed in this work. Of these, thirty-four were resolved well enough for the determination of their relative concentrations. Figure 5.4 shows the location of the 34 resolved components, where the numbers 1-18 refer to the peaks initially detected upon visual inspection of the data, and number N1-N16 refer to the newly detected peaks. The indicated subsection of the chromatogram shown in Figure 5.4 and Figure 5.5A was used to quantify peak N16. The spectral and chromatographic profiles obtained after implementation of IKSFA-ALS-ssel are shown in Figure 5.5B. As can be seen in this figure, the IKSFA-ALS-ssel analysis revealed the presence of eight components in this subsection. Two of these components were identified as background components. It should be noted that the analysis of additional overlapping

subsections in this region of the chromatogram permitted peaks 10 and N8, and peaks 9 and N15 to be resolved from one another, as well as resolving peaks N10, N11 and N12 from the background for a total of ten quantified peaks. While the above-mentioned peaks have the same first dimension retention times and very similar second dimension retention times, chemometric

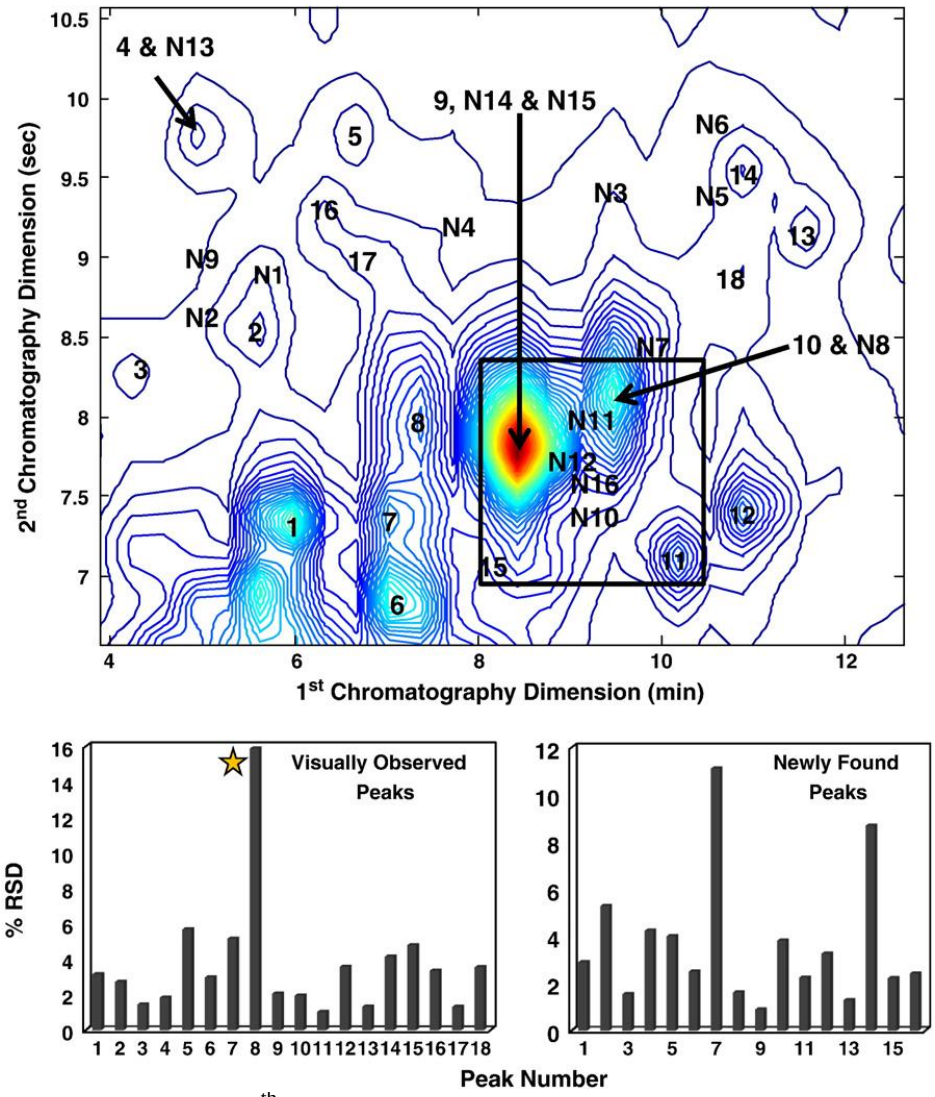

Figure 5.4: Contour plot of the 7th urine control at 216 nm showing 34 resolved peaks. The N preceding 16 of the 34 resolved and quantified peaks signifies that those peaks were found and resolved only after application of the developed chemometric method (newly found) while the other 18 peaks were visually observable prior to chemometric analysis. The two bar graphs show % RSD values calculated for the corresponding peaks. The star on the visually observed peaks graph indicates that Peak 8 is considered to be a chemically unstable compound.

resolution was possible due to the unique spectra of the corresponding peaks. The ability of the algorithm to resolve chromatographically overlapped peaks having different spectral profiles was demonstrated in several areas of the data in which two or more peaks were found to be present, but only one peak was visually apparent. Evidence of several additional very weak peaks was also found in this subsection, but these peaks could not be reliably quantified.

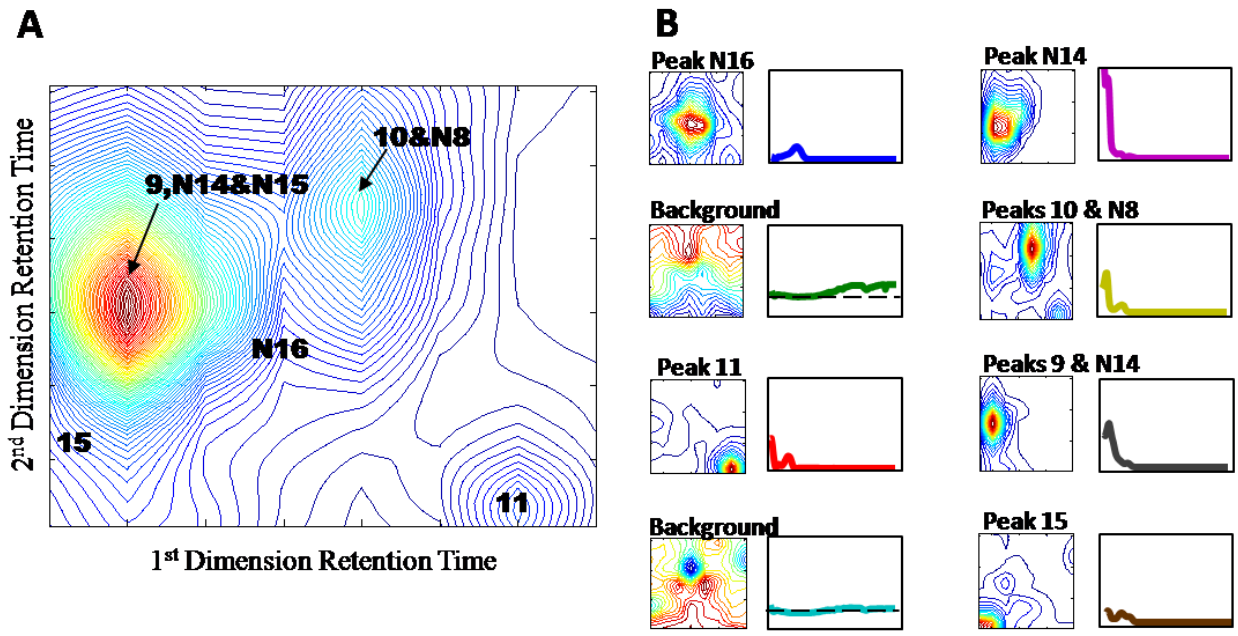

Figure 5.5: (A) Contour plot of a subsection of urine control data at 216 nm in which resolution and quantification of Peak N16 is the goal for the chemometric analysis. (B) The chromatographic and corresponding spectral results for each component of the 8 component IKSFA-ALS-ssel analysis for the above subsection of raw data.

The bar graph in Figure 5.4 provides the % RSD values determined for the chemometrically resolved peaks. The % RSD values for the initially observed peaks ranged from 1.04 % for peak 11 to 15.9 % for peak 8, with an average % RSD of 3.73 %. Peak 8 appeared to be a chemically unstable compound (its intensity consistently decreased over the course of the analysis) leading to the poor quantitative precision for that compound. The % RSD values for the sixteen additionally found peaks ranged from 0.90 % for peak N10 to 11.1 % for peak N7

with an average % RSD value of 3.56 % for this group of resolved peaks. This section for all fourteen replicate injections was also evaluated by LCImage software [22] using their blob detection tool (*i.e.,* peak picking). The blob detection found on average 22 peaks using the default settings and 24 peaks after modification of the detection setting in the section of data analyzed in this work. The number of detected peaks for the fourteen replicate injections ranged from 20 to 28 depending on the injection and the setting used for detection. Of the 24 peaks found by the LCImage software for sample injection 7, three were also found by the IKSFA-ALS-ssel method but are cut off by the section parameters and therefore not included in the 34 quantified peaks. Also, two of the LCImage detected peaks for injection 7 are not detected in all fourteen injections.

The relative signal was evaluated and compared to the corresponding % RSD values for each quantitatively resolved peak to determine if a low signal response was correlated to a decrease in the precision of quantification as seen by an increase in % RSD values. The relative signal response was determined by multiplying the chromatographic maximum value of the 7^{th} sample injection by the spectral maximum value for each peak. This assumption can be made due to a relatively constant background response of the section of the data analyzed. We found that in the majority of cases where the % RSD of a given peak is above 4, a low signal response was not responsible for the observed poor precision, but rather other chromatographic phenomena such as spectral or chromatographic rank deficiencies (overlapped peaks) and unsatisfactory resolution of the peaks from the background. These issues will be discussed in more detail in Chapter 7. Peaks with % RSD values of less than 2 % were not affected by these issues.

5.5 Comparison to Previous Rutan Group Work

Previously, Porter *et al.* analyzed four-way data arising from a comprehensive LC × LC analysis of maize seedlings [2]. Retention time shifting was thought to be minimal or nonexistent, as the total run time for all samples was only three hours, the data was assumed to be approximately quadrilinear. This assumption allowed Porter *et al.* to employ the PARAFAC model such that the results from PARAFAC were used to initiate the in-house ALS algorithm so that constraints could be applied selectively. The samples used for method comparison in this work consisted of two extracts of mutant orange pericarp maize seedlings and two extracts of wild-type maize seedlings.

For method comparison purposes, a small section of the previously analyzed data set, shown in Figure 5.6, was analyzed using the current IKSFA-ALS method with the exception that the spectral selectivity constraint was not employed due to insufficient wavelength collection

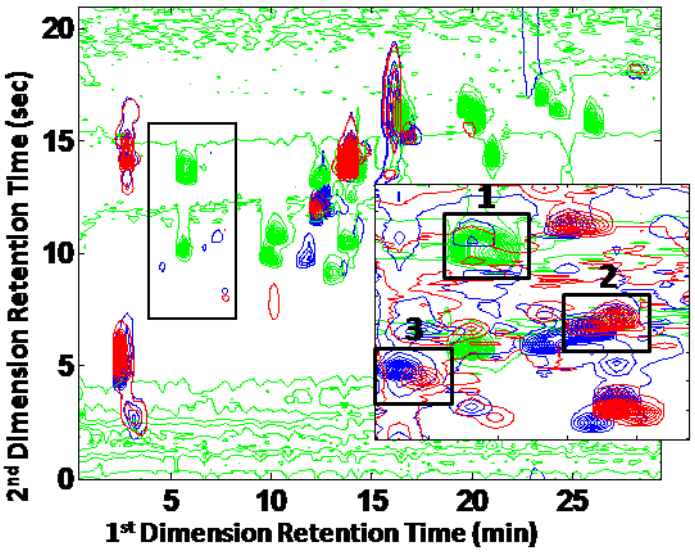

Figure 5.6: Overlaid contour plot of maize data analyzed by Porter *et al.* [2]. The blue contour plot is the first injection of the mutant sample, the green contour plot is the indole standard mixture and the red contour plot in the second injection of the wild-type sample. The inset corresponds to the outlined section.

during the LC × LC run of the maize data. Peaks labeled 1, 2, and 3 indicated by the boxed subsections within the inset of Figure 5.6 were chosen for % RSD comparison of the determined relative concentrations. These peaks were selected because they were present in both the mutant and wild type samples and were resolved using the PARAFAC- ALS method. The results in Table 5.3 show that for four out of the five comparisons made, IKSFA-ALS yields considerably lower % RSD values for these peaks than PARAFAC-ALS. It is of particular interest, that the IKSFA-ALS method chemometrically resolved an additional six peaks that were not detected by Porter *et al.* and was able to resolve several peaks that the PARAFAC-ALS method did not resolve due to the lack of multilinearity in the first retention time dimension. We also conclude from these results that, even for the relatively short three hour run, there were sufficient retention time shifts to decrease the precision of the PARAFAC-ALS analysis, relative to the IKSFA-ALS method.

Table 5.3: Comparison of % RSD values for duplicate samples resulting from PARAFAC-ALS method [2] and IKSFA-ALS methods in the analysis of maize data. NP: the compound was not present in the wild-type samples.

	PEAK 1		**PEAK 2**		**PEAK 3**	
	PARAFAC-ALS	IKSFA-ALS-ssel	PARAFAC-ALS	IKSFA-ALS	PARAFAC-ALS	IKSFA-ALS
Mutant	**40.1**	**5.4**	**141.0**	**14.6**	**5.1**	**1.4**
Wild Type	NP	NP	**26.4**	**2.5**	**21.8**	**82.0**

5.6 Data Analysis Considerations

Due to the size of the urine control data set and to the large number of factors involved in the analysis of an entire chromatogram, it was first necessary to divide the data into sections. This enabled us to work with a more manageably sized section shown in Figure 5.3B from 3.85 to 12.6 minutes and 6.6 to 10.6 seconds; this section was chosen for further investigation since it is free of signals where the detector was saturated. The nature of the data (complex and lacking

multilinear behavior because of the retention time shifts) limits the chemometric methods available, requiring that either prealignment data processing occurs before chemometric implementation of methods that require multilinearity can be employed, such as PARAFAC, or restriction of the data analysis to methods that are not affected by retention time shifting such as MCR-ALS. We chose the second option, employing an approach involving IKSFA and MCR-ALS, neither of which requires multilinearity. The authors recognize that the described method requires user intervention. While the number of components, *N*, determination step and the spectral selectivity constraints implementation require the user to make decisions based on visual inspection of the results before proceeding to the next step, these decisions are fairly straightforward and are not time consuming. In other words, this method can be easily taught and learned such that a great deal of expertise is not required to achieve good results. In addition, the method is shown in section 5.5 to be applicable to other data sets arising from LC × LC-DAD analysis and to be an improvement over a previously published method.

While the number of components (*N*) is somewhat subjectively determined, it is easily and quickly accomplished. Contour subplots of the subsection to be analyzed at different wavelengths allows for an approximate number of peak components to be determined by simply counting the peaks that are visually apparent. Due in large part to the large dynamic range of this data, chromatographic peaks are not always observable even when plotted at multiple key wavelengths. Hence, the comparison of the number of visually counted peaks to the number of principal components ascertained from the scree plot leads to a reasonable initial estimate of *N* that can be attained in less than a minute. It is important to keep in mind that there are also background components to be considered to obtain the final estimate for the number of spectrally distinct components, *N*. The determination of an appropriate final *N* parameter included the

consideration of several factors: there should be no more than 3 background components following curve resolution, the value of the determinant for the final key set should be less than 0.1 and greater than 0, and the fit error for the MCR-ALS step should be less than 5%. Addition of more components to reduce the fit error further usually resulted in overfitting, as evidenced by the appearance of the component profiles that did not make sense chromatographically or spectrally. A cross validation of the subsection used for the analysis of peak N16 in which a leave-seven-out approach was taken for component models of N= to 7, 8, and 9, confirmed that for this subsection the eight component model chosen using the above described method resulted in the best fitting model.

The second manual step that we employ is the implementation of the constraints. One of the advantages of the in-house MCR-ALS algorithm [60] is that each of the constraints can be selectively applied to individual components. An example of this is the selective application of the spectral selectivity constraint to only the non-background components such that the wavelengths from 440 nm to 700 nm were set to zero. This wavelength range was chosen due to the complete lack of corresponding spectral information above 440 nm to any components other than the background. This helps the algorithm resolve the background from actual components because the background spectra have a consistent increasing absorbance above 440 nm. Also, the manner in which the spectral selectivity constraint was employed, allowed for the selective application of the non-negativity constraint to the spectral dimension of all components except the background components. The implementation of the two spectral constraints aids in the spectral resolution of the background spectral components from non-background spectral components, as illustrated in Figure 5.7. Chemometric resolution of the background components

provides substantial improvements in quantification using the manual baseline method for relative concentration determination.

The unimodality constraint was not employed in this work for several reasons. Unimodality, as is currently employed in many MCR algorithms, sets a vertical at the valley of the non-unimodal peak and sets all of the data points of the peak with the smaller maximum to zero [11, 60, 103]. This, in essence, eliminates a possible smaller peak from the analysis results that the manual baseline method may be capable of integrating. Alternatively, dynamic unimodal regression may be used; however, in practice the smaller peak is still lost [104]. It is important to remember that an incompletely resolved component may be non-unimodal in either

the first dimension retention time, the second dimension retention time or it may exhibit non-unimodal behavior in both retention time dimensions. What would ultimately be required is the capability to employ the unimodality constraint for four-way data to selective components in a manner that adds an additional component to the result and assigns the smaller of the non-

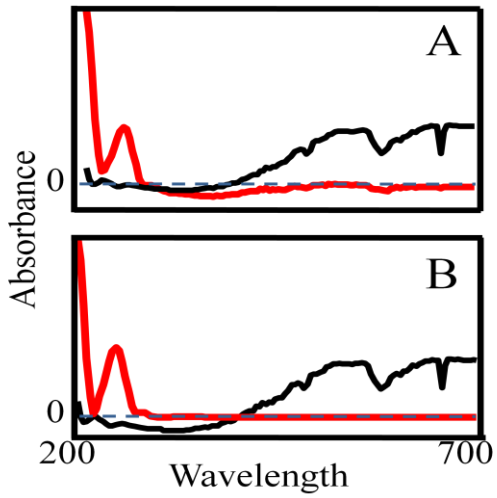

Figure 5.7: Component spectra (black) and background spectra (red) before (A) and after (B) implementation of the spectral selectivity and spectral non-negativity constraints. By zeroing the chemical component spectra after 440 nm, the algorithm is better able to resolve the background spectra from the compound spectra.

unimodal peaks to the "new" component so that no information is lost in the resulting answer. Based on the motivation provided by this work, such an algorithm has now been developed [105]

5.7 Conclusions

Most of the published peak detection methods for LC × LC data analysis [22, 46, 106, 107] have been for chromatographically well-resolved peaks. Curve resolution procedures that have been useful for the analysis GC × GC data [16, 108] for the most part have not been successfully applied to LC × LC, probably because of the same retention time reproducibility issues that we encountered in this work. Also, the modification of successful algorithms used for the analysis of 1D techniques to 2D chromatography is complicated due to the under sampling effect of the first dimension and the necessity of combining several second dimension peaks to represent the total LC × LC peak [49, 106]. We have shown that the IKSFA-ALS-ssel method successfully resolves complex LC × LC-DAD data without requiring prealignment of the data to achieve multilinearity. Due to lack of retention time alignment, the previously developed PARAFC-ALS method showed higher % RSD values, assigned the same peak to different components and did not resolve peaks that were found to be present when compared to the IKSFA-ALS method. The current drawback to the IKSFA-ALS-ssel method is the lack of automation. However, the intervention that is required is straightforward and relatively simple, if somewhat tedious. For the standards mixture data, there is a 2.5 fold improvement in the % RSD values of the IKSFA-ALS-ssel analyzed data as compared to the raw data. The chemometric analysis of the urine control data revealed over fifty compounds, thirty-four of which were resolved sufficiently for quantitative analysis. The average % RSD of the quantified peaks of 3.5 %, while rather high for accepted 1D-LC analysis, is quite good for such a complex

sample such as human urine but leaves room for improvements in quantification of LC × LC data.

Several issues associated with the quantification of this data arose during curve resolution, including phase shifting caused by retention time shifts in the first dimension, rank deficiency, large dynamic range issues and unsatisfactory curve resolution of the peaks from the background. These are several of the obstacles associated with achieving more precise quantification and are addressed in Chapter 7.

Chapter 6: Chemometric Analysis of Targeted 3DLC-DAD Data for Accurate and Precise Quantification of Phenytoin in Wastewater Samples

Adapted from H.P. Bailey, S.C. Rutan, D.R. Stoll, Journal of Separation Science, 2012, 35, 1837-1843

A variety of pharmaceuticals have been found in various water systems, including wastewater treatment effluent. Due to the possible environmental and human health implications, it is important to be able to quickly and reliably quantify the amount of pharmaceuticals and personal care products that may be present in such samples. To this end, a new chromatographic analysis technique involving three dimensions of liquid chromatography, including selective comprehensive separations in the second and third dimensions, was applied to the analysis of a wastewater treatment plant effluent (WWTPE) sample using both standard addition and external calibrations. Iterative key set factor analysis alternating least squares with the application of both sample and spectral selectivity constraints was used to resolve the phenytoin peak at a concentration corresponding to about 40 parts-per trillion using UV absorbance detection. Both the precision and accuracy of the method are investigated in this chapter.

Stoll *et al.* have developed a novel LC approach, coined selective comprehensive two-dimensional HPLC (sLC × LC), such that the s stands for the selective heartcutting of the ^{1}D separation to include the analyte of interest. This approach was shown to combine the advantages of heartcutting two-dimensional LC (LC-LC) and LC × LC, while eliminating the

disadvantage arising from "the long standing link between the timescales of the ^{1}D and ^{2}D separations of conventional online LC × LC which "preserves the ^{1}D resolution of one or more target compounds from closely neighboring peaks" [101]. Data arising from the 3D-LC separations of water extracts, where sLC × LC was used in the second and third dimensions were analyzed using the chemometric approach iterative key set factor analysis-alternating least squares with spectral selectivity (IKSFA-ALS-ssel). Chapter 5 discussed the implementation of the developed IKSFA-ALS-ssel to urine control samples arising from a LC × LC separation. In that analysis, accuracy of the chemometric method was not reported, only precisions for the fourteen replicate injections of a standard urine control sample, owing to the lack of a calibration set [1, 13].

The nature of the phenytoin dataset allows for the characterization of both the precision and accuracy of the chemometric methodology. The procedure for the analysis of this data follows the previously published IKSFA-ALS-ssel method [1] described in Chapter 6 but with a few significant modifications necessary to accommodate some features of the phenytoin dataset. The first modification involved a change in the spectral selectivity constraint (refer to Figure 3.8 and section 3.4 for the general discussion of this constraint). The range of the spectral selectivity constraint was modified in this work for two reasons. It allowed for accommodation of the decrease in the total number of wavelengths analyzed. Second, it was determined that the peaks of interest did not exhibit any spectral response above 360 nm while the background signal had a distinct response at these higher wavelengths due to changes in refractive index, thus increasing the ability of the algorithm to better distinguish the background signal from the signals of interest. This refractive index effect has three sources: (1) the mismatch of the first dimension mobile phase containing a percentage of organic modifier with that of the initial 100 % aqueous

mobile phase of the second dimension, (2) the rapid gradient of the second dimension and (3) rapid re-equilibration of the second dimension column returning to 100 % aqueous. The second, and most significant modification, was the implementation of an additional constraint for sample selectivity [109]. This constraint was used to compensate for the lack of an interferent peak in the DI water samples and is further discussed in section 6.2. To the author's knowledge, this constraint has not been previously employed in the chemometric analysis of 3D-LC data.

Prior to chemometric analysis, the dataset was sectioned to encompass only the region containing the phenytoin and interferent peaks, as shown in Figure 6.1 (details were provided in section 4.3) where the shaded portions of the contour plots were eliminated from the data analysis. By analyzing a limited section of the data, the background signal is minimized, and this allows for the best resolution of the two compounds from the background signal. This concept, the effects of the size of the analysis window on resolution and quantification, was previously explored in the work of section 5.3.

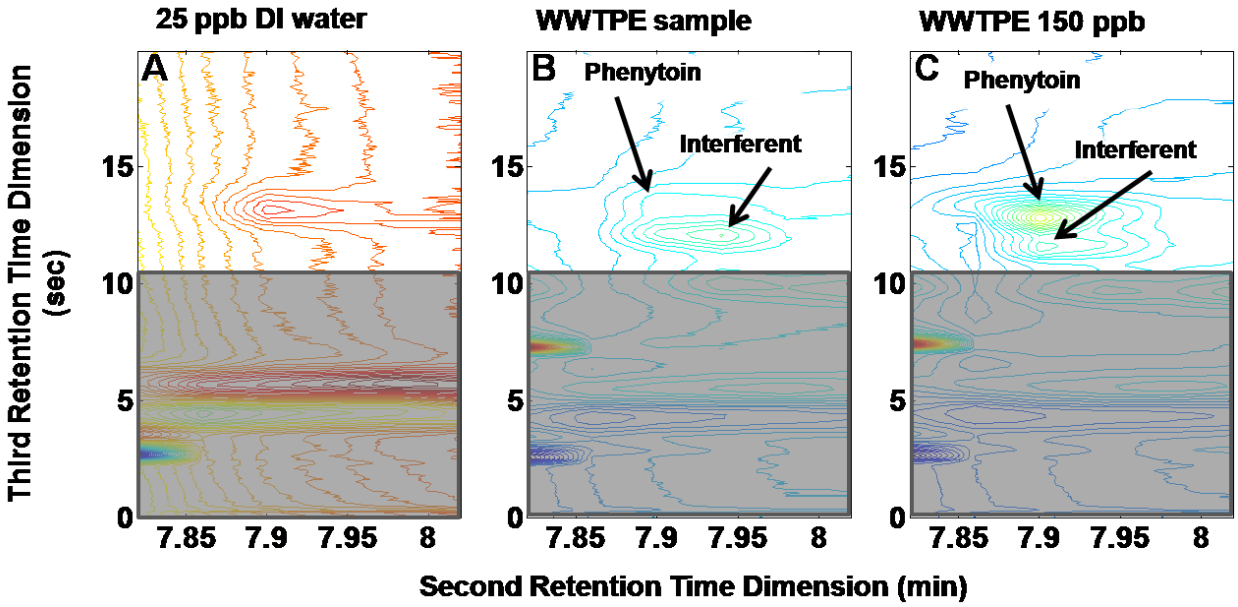

Figure 6.1: Contour plots of various sample injections at 216 nm before chemometric analysis. The shaded portion of the plots is the section of the data eliminated from the chemometric analysis of the data. (A) Contour plot of DI water sample spiked with 25 ppb phenytoin. (B) Contour plot of the WWTPE sample without a spiked amount of phenytoin. (C) Contour plot of the WWTPE sample spiked with 150 ppb phenytoin.

6.1 IKSFA-ALS-ssel

A six component model was used for the IKSFA-ALS-ssel analysis. The chemometric results are shown in Figure 6.2A where the chromatographic profiles of the six component IKSFA-ALS-ssel model for the 75 ppb phenytoin sample (in DI water) and the WWTPE sample spiked with 75 ppb of phenytoin are shown in the first and second rows of the contour plots, respectively. Components 1, 3, 4 and 6 are associated with the background signal and the matrix signal, component 2 in the figure shows the phenytoin peak, which has its own corresponding spectrum and component 5 is associated with the interferent when resolved properly.

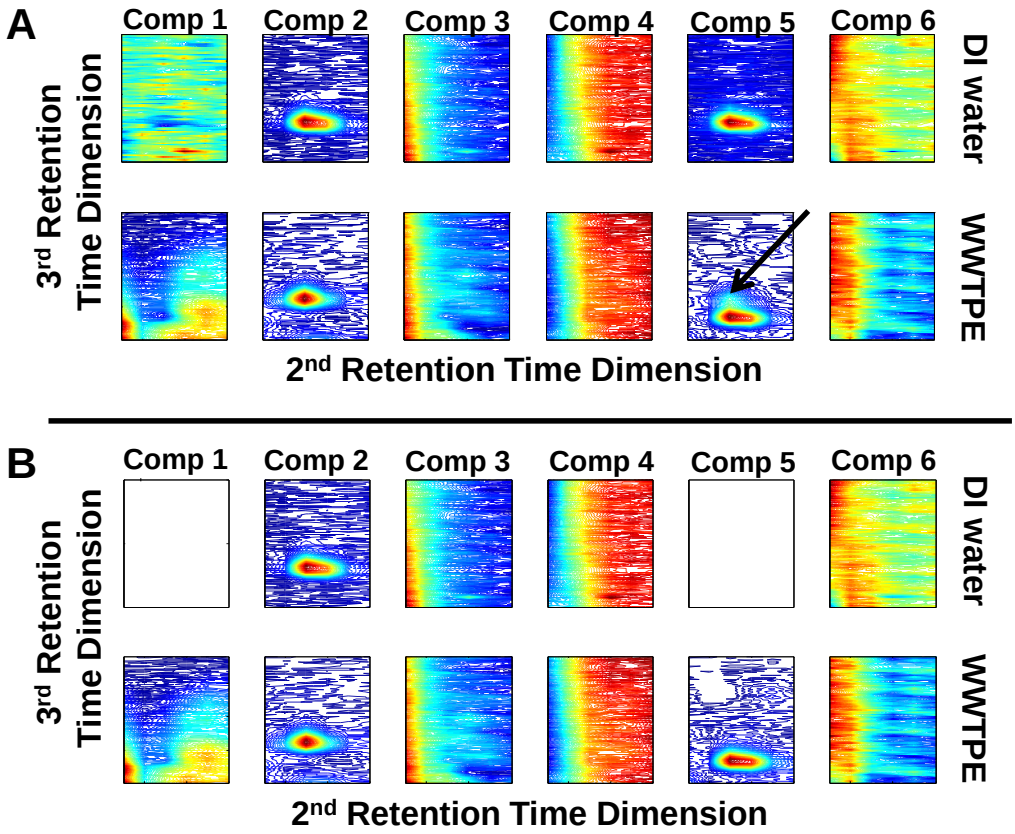

Figure 6.2: Chromatographic results of the chemometric analysis for a six component model for the 75 ppb phenytoin standard sample and the 75 ppb phenytoin in addition to the WWTPE sample. (A) Analysis without implementation of the sample selectivity constraint which overfits the DI water samples and assigns some of the phenytoin peak to component 5 (as indicated by the arrow), the interferent component in the WWTPE samples. (B) Analysis with implementation of the sample selectivity constraint such that the concentrations of components 1 and 5 of the DI water samples were constrained to be zero.

An investigation into component 2 (phenytoin) and component 5 (interferent) of the WWTPE sample injections after IKSFA-ALS-ssel analysis, reveals an incomplete resolution of the analyte and the interferent as indicated by the arrow in Figure 6.2A. This chromatographic contour plot for component 5 for the 150 ppb spike of the WWTPE sample is expanded in Figure 6.3A. The maximum of the interferent peak is located at 11.50 seconds and 7.9 minutes of the 3rd retention time dimension and 2nd retention time dimension axes, respectively; while the maximum of the phenytoin peak is located at 12.85 seconds and 7.9 minutes, 3nd retention time dimension and

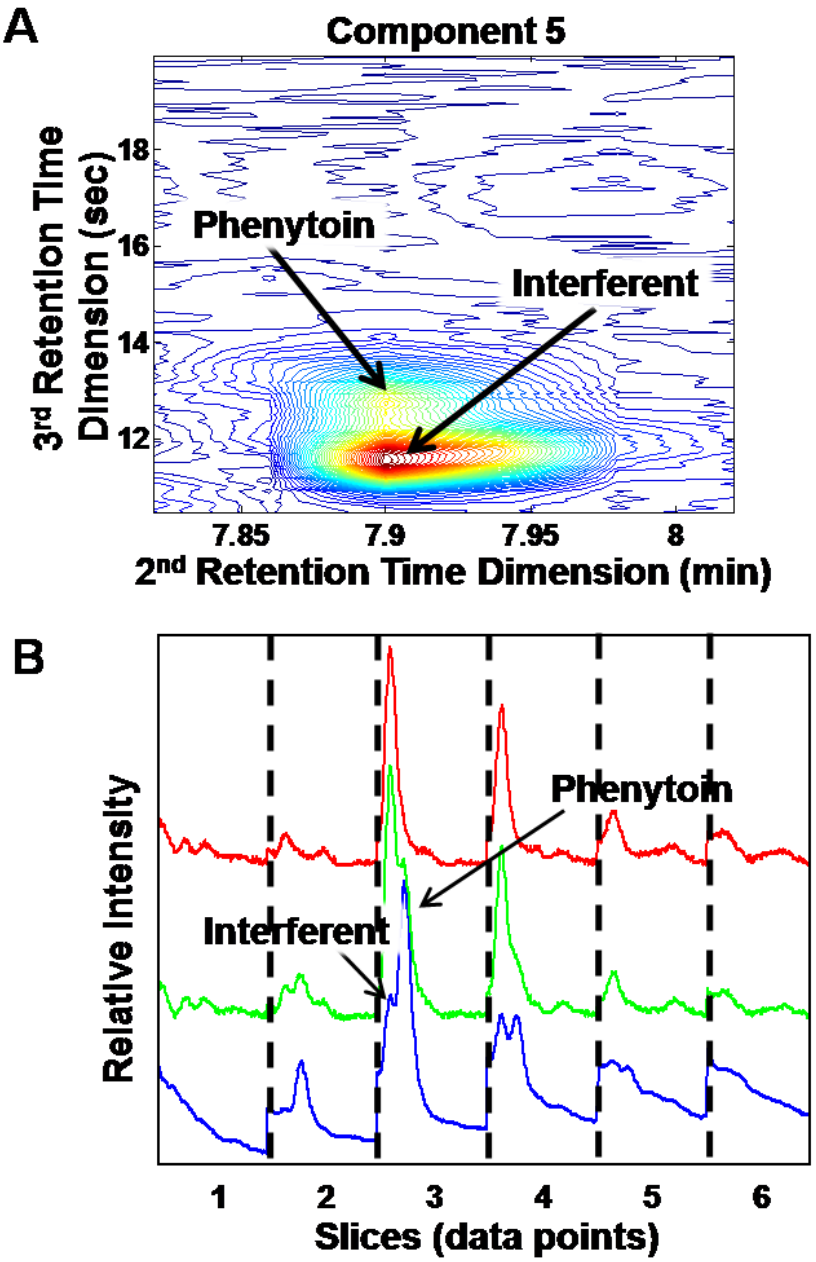

Figure 6.3: Plots showing the overlap of the phenytoin and interferent peaks. (A) Contour plot of the fifth IKSFA-ALS-ssel component for the 150 ppb spiked WWTPE sample which shows the incorrect assignment of a portion of the phenytoin peak eluting after the interferent peak in the third retention time dimension. (B) Overlay of the 150 ppb spiked WWTPE sample for three third dimension sequenced chromatograms, such that the blue (bottom) series of chromatograms is the raw data, the green (middle) series of chromatograms is the IKSFA-ALS-ssel analyzed data for the fifth component, which shows the incomplete resolution of the phenytoin and interferent peaks and the red (top) series of chromatograms is the IKSFA-ALS-ssel-csel result for the fifth component, which shows complete resolution of the phenytoin from the interferent peak.

2nd retention time dimension axes. This incomplete resolution leads to poor accuracy and precision of quantification of the analyte. Table 6.1 shows the poor precision of the phenytoin peak with the RSDs of the duplicate injections ranging from 0.38 % for the DI water at 150 ppb to 7.6 % for the WWTPE at 25 ppb.

Table 6.1: % RSD of the duplicate sample injections for both the DI water and WWTPE samples after chemometric analysis

Spiked conc (ppb)	ssel % RSD	ssel-csel % RSD	2D-LC-MS % RSD
DI 25	6.3	0.19	
DI 50	6.7	0.34	
DI 75	4.6	1.2	
DI 150	0.38	0.32	
WWTPE	4.1	0.65	25
WWTPE 25	7.6	0.37	20
WWTPE 50	5.2	0.40	17
WWTPE 75	5.7	3.1	8.8
WWTPE 150	1.9	0.45	8.9
AVERAGE	4.7	0.77	15

a)ssel: implementation of spectral selectivity only.
b)ssel-csel: implementation of both the spectral and sample selectivity constraints.

6.2 IKSFA-ALS with all constraints

To solve the overfitting of the DI water samples and the incomplete resolution of the interferent and phenytoin peaks in the WWTPE samples, the sample selectivity constraint was employed for the first time for this type of separations data. Our in-house ALS algorithm allows for the selective application of constraints. In other words, any constraint can be imposed on a given data dimension or on all data dimensions; and within a given dimension, the constraints can be imposed on a select few or all components. This allows for the flexible application of the constraints to more accurately represent the known chemistry of the samples, thereby aiding in the elimination of mathematically correct solutions that are not chemically valid. To this end,

the sample selectivity constraint was applied in a twofold manner: by component and by sample injection. Hence, only those components determined to be overfit in the DI water samples, but were required to appropriately fit the WWTPE samples, were constrained. While the background components 3, 4 and 6 are consistent for both the DI water samples and the WWTPE samples, a comparison of component 1 in Figure 6.2A shows that this component is different for the two different sample types; thus component 1 (an overfit background) and component 5 (the overfit interferent) are constrained. The second implementation of the sample selectivity constraint was to the sample injections of phenytoin in DI water, where no interferent peak was present, *i.e.*, sample injections 1-10. The selected components (components 1 and 5) and injections (injections 1-10) for constraint were set to zero and the results are shown in Figure 6.2B.

There are two significant points of interest with respect to the implementation of the IKSFA-ALS algorithm on this data set. The use of the sample selectivity constraint on the 1^{st} and 5^{th} component of the DI water samples forces the entire phenytoin peak appropriately into component 2. Also, complete resolution of the phenytoin peak (component 2) from that of the interferent peak (component 5) in the WWTPE samples is also achieved. This greatly improved resolution using the sample selectivity constraint is shown in Figure 6.3B. The bottom series of ^{3}D chromatograms (blue) are from the WWTPE sample spiked with 150 ppb phenytoin before any chemometric data analysis was performed on the dataset. In this sample the phenytoin gives a larger signal than the interferent so that the interferent appears as a shoulder to the left of the major peak, phenytoin. From this chromatogram, it is clear that accurate quantification of phenytoin is not possible due to the severe overlap seen in each of the ^{3}D chromatograms of the raw data sequenced chromatogram. The middle series of chromatograms (green) in Figure 6.3B

is the IKSFA-ALS-ssel analysis (without the sample selectivity constraint) result for component 5 (the interferent peak) and clearly shows that the phenytoin is present as a shoulder to the right of the major interferent peak since it elutes after the interferent in the ^{3}D. This leads to the poor precision of the method as described in section 6.1. However, the top chromatogram (red) is the chromatographic result for component 5 after resolution with the sample selectivity constraint. From this, it is clear that the sample selectivity constraint has resolved the interferent from the phenytoin peak such that none of the phenytoin is being inappropriately assigned to the interferent component. This leads to better precision and accuracy of quantification for the unknown samples. RSDs for the duplicate injections range from 3.11 % for the WWTPE at 75 ppb and 0.19 % for the DI water sample at 25 ppb, as shown in Table 6.1. The average % RSD of the duplicate injections is 0.77 after implementation of the sample selectivity constraint, as shown in Table 6.1, which is a 6-fold improvement over the average precision of the duplicate injections observed without the implementation of the sample selectivity constraint.

6.3 Statistical Analysis

This particular dataset allows for determination of phenytoin concentration of the unspiked sample using either the standard addition method (*i.e.*, using the spiked series of WWTPE samples), or the external calibration method (*i.e.*, using the spiked DI water samples as calibrants). The concentration of phenytoin in the unknown sample (without the sample selectivity constraint) was determined to be 32 ± 3 ng/L using the standard addition method and 31 ± 4 ng/L using the external calibration method, as shown in Table 6.2. A 2D-LC-MS/MS analysis for the same WWTPE sample resulted in a phenytoin concentration of 43 ± 5 ppb (all error estimates given as standard errors) [101]. The low calculated phenytoin concentration resulting from the standard addition method is directly related to the incomplete resolution of the

interferent peak since a portion of the phenytoin is incorrectly assigned to component 5 for all sample analysis, that portion cannot be quantified.

The concentration of phenytoin in the unspiked WWTPE using both the spectral and sample selectivity constraints was determined to be 42 ± 1 ng/L using the standard addition method and 36 ± 1 ng/L using the external calibration method, as shown in Table 6.2. This is an improvement in accuracy and precision over the chemometric results obtained without the sample selectivity constraint, and agrees with the results achieved using the 2D-LC-MS/MS method. The slightly lower results for the external calibration method are likely due to matrix effects in the WWTPE samples. To test this, we compared the slopes of the regression lines from both methods [110, 111]. An F-test was first done to determine if the variances of the two slopes were statistically different, followed by a t-test. The confidence interval for the difference in the two slopes was determined to be 0.012 ± 0.003 with a probability of the two slopes being statistically similar of $p < 0.0001$, *i.e.*, the slopes are different; and therefore, a matrix effect is present.

Table 6.2: Comparison of the unknown sample calculations using both the standard addition and calibration methods for the chemometric method with and without the sample selectivity constraint ssel constraint

	ssel constraint		**ssel and csel constraint**	
	concentration	**s_y**	**concentration**	**s_y**
Standard Addition[a]	32 ± 3 ng/L[b]	0.397	42 ± 1 ng/L[c]	0.990
Calibration Method	30 ± 4 ng/L[d]	0.467	36 ± 1 ng/L[e]	0.139

a) 2D-LC-MS result 43 ± 5 ng/L, y = 0.025 (± 0.002) x +1.1 (± 0.1), $n = 10$, $R^2 = 0.96$, $s_y = 0.29$.
b) y = 0.082 (± 0.002) x +2.6 (± 0.2), $n = 10$, $R^2 = 0.992$.
c) y = 0.078 (± 0.001) x +3.28 (± 0.09), $n = 10$, $R^2 = 0.998$.
d) y = 0.088 (± 0.003) x−0.3 (± 0.2), $n = 10$, $R^2 = 0.992$.
e) y = 0.0901 (± 0.0009) x −0.16 (± 0.07), $n = 10$, $R^2 = 0.999$.

We completed a statistical analysis of two comparisons: (1) the concentration derived from the standard addition curve for the 2D-LC-MS/MS method and for the IKSFA-ALS method

using both constraints, and (2) the concentration derived from the standard addition curves for IKSFA-ALS method using only the spectral selectivity constraint and for IKSFA-ALS method using both the spectral and sample selectivity constraints. The standard deviations of the calculated concentrations of the 2D-LC-MS/MS and the IKSFA-ALS method using both constraints were found to be statistically different ($p = 0.00015$ probability of being incorrect in saying that the variances are different), thus requiring the use of the unequal variance t-test for the comparison of the derived concentrations. As expected, there was no significant difference between the sLC × LC-DAD (42 ± 1 ng/L) and 2D-LC-MS/MS (43 ± 5 ng/L) estimated concentrations of phenytoin in wastewater ($p = 0.91$). However, there was a significant difference ($p = 0.012$) in the IKSFA-ALS results upon implementation of the sample selectivity constraint. When only the spectral selectivity constraint was employed a concentration of 32 ± 3 ng/L was found *vs.* 42 ± 1 ng/L when using both the spectral selectivity and sample selectivity constraints.

6.4 sLC × LC Importance

At this point we call attention to the significance of the sLC × LC approach to the resolution of phenytoin in the presence of the unknown interferent in the WWTPE sample. Prior to our initial analysis of this sample we had no way of knowing that there would be a major interferent peak overlapping the phenytoin peak in the second dimension time axis, but it turns out that this particular arrangement of peaks provides the opportunity to highlight the advantage of the sLC × LC approach compared to either heartcutting or LC × LC analyses of the same sample. The apices for the phenytoin and interferent peaks are slightly offset in the ^{2}D retention axis. This small offset is important to the success of the multi-way analysis algorithm. In contrast to the sLC × LC approach where the phenytoin/interferent peak is sampled frequently,

this slight separation would be completely lost in both the heartcutting and LC × LC cases due to much larger sampling times (*i.e.*, relative to the ^{2}D peak width), and would reduce the likelihood that the IKSFA-ALS algorithm would be able to resolve the two peaks to the level needed for accurate and precise quantitation of the phenytoin target compound.

6.5 Conclusions

Due to the severe chromatographic overlap of the phenytoin and interferent peaks chromatograms resulting from multi-dimensional separation of WWTPE samples, accurate quantification was not possible without sophisticated data treatment. The 3D-LC selective comprehensive separation in conjunction with the chemometric analysis using a sample selectivity constraint provides enough resolution from the many other compounds in the WWTPE sample for successful quantification. Thus, the need for further chromatographic method development is negated. This analysis of replicate spiked DI water and WWTPE samples allowed for both accurate and precise determination of phenytoin in WWTPE, as well as an evaluation of the accuracy and precision of the IKSFA-ALS method described in Chapter 5. We have shown that there is a considerable improvement in both precision and accuracy in phenytoin quantification when the sample selectivity constraint is applied to the DI water samples. This is due to the complete resolution of the overlapped peaks and to the correct component assignment of these two peaks upon implementation of this constraint. The average precision of the duplicate phenytoin measurements after implementation of constraints was improved by a factor of twenty compared to previously published 2D-LC-MS/MS results. The results of the analysis without the sample selectivity constraint were found to be significantly different from that of the analysis with both the spectral selectivity and sample selectivity

constraints; there was excellent agreement between the 2D-LC-MS/MS method and the IKSFA-ALS after implementation of both the spectral and sample selectivity constraints.

Chapter 7: Factors that Affect Quantification of Diode Array Data in Comprehensive Two-Dimensional Liquid Chromatography using Chemometric Data Analysis

Adapted from H.P. Bailey, S.C. Rutan, J. Chromatogr. A, 1218 (2011) 8411-8422

To date, the central analytical issue relevant to LC × LC separations, quantification, has received only minimal attention. It is vital to the further development of this technique that a greater understanding of the specific factors affecting peak quantification of LC × LC be attained. In the vast majority of the reports, only well-resolved peaks were quantified (see Chapters 2 and 5). However, for the quantification of large data sets, before anything resembling "ideal conditions" (well resolved peaks) can be achieved, it is first essential to resolve the overlapped peaks. The chemometric resolution can be complicated because the data arising from LC × LC analysis of complex samples typically consist of multiple compounds that elute at very similar retention times and of multiple compounds that have the same or very similar spectra. As described in Chapter 5, we developed a curve resolution method for the resolution and quantification of LC × LC data [1]. The results from that study allowed us to investigate in more detail several key issues that affect peak quantification in LC × LC-DAD data. These issues are the subject of this chapter and include data size (124.5 million data points –approximately 1 GB), spectral and chromatographic overlap, retention time shifts, dynamic range issues and inadequate removal of the background signal from the data. An understanding of these issues and their

effects on peak quantification is critical for the application of LC × LC methods for the quantification of analytes present in complex mixtures.

This chapter explores the impact of these issues on the effectiveness of LC × LC as a technique for the quantitative analysis of complex samples. The above mentioned factors that affect peak quantification are investigated using fourteen replicate analyses of a urine sample (see section 4.1 for details on the data analyzed), representing the effects of such factors when analyzing samples in complex matrices. We demonstrate that quantification of LC × LC data is improved following implementation of chemometric techniques that minimized the deleterious effects to quantification due to chromatographically overlapped peaks, retention time shifting and background signal interference. The chemometrically resolved data shows a 2.5-fold increase in precision of quantification over the quantification of the raw data. It is also demonstrated that the method quantifies sixteen peaks that were not visually present prior to chemometric analysis.

7.1 Review of the Implemented Chemometric Method

Briefly, a section of the data where the absorbance was less than two was chosen for chemometric analysis; and due to the complexity and size of the data section, the data were further divided into subsections. The next step was to determine the number of components (unique spectra) in the data subsection to be analyzed, followed by resolution of the chromatographic peaks using an in-house MCR-ALS algorithm [60]. Each unique spectrum is assigned to an individual component and all chromatographic peaks (compounds) associated with that spectrum are also assigned to that component. After application of the IKSFA-ALS-ssel algorithm to obtain resolved peak profiles, relative peak signals were calculated by manually integrating each second dimension peak and summing the areas. This manual baseline method

was previously described [1, 50]. In short, a sequence of second dimension chromatograms for a given peak is plotted and a baseline is manually drawn for each second dimension peak. The areas for the peaks are determined and the volume of the corresponding first dimension peak is calculated by simply summing the areas of the second dimension peaks. Peak quantification was also carried out using LCImage software v 2.1 (GC Image, LLC Lincoln, NE) [22]. The default parameters were used for baseline correction and volume determinations. Percent RSD values (determined by dividing the standard deviation of the calculated volumes for a given species for all sample injections by the average peak volume and multiplying by 100) were calculated for the resolved peaks in both the replicate standard mixture samples and in the urine control samples.

7.2 Comparison of Quantification Methods

Several different methods of peak size determination for raw chromatograms and chemometrically analyzed data were compared based on the six replicate injections (see Table 7.1) of the standard mixture. LCImage software and manual baseline methods were used for quantification of the raw data. A simple summation method, LCImage software and a manual baseline method were utilized for quantification of the IKSFA-ALS-ssel analyzed data. The summation method simply adds all the intensities of the reconstructed chromatogram corresponding to a given spectral component within the subsection. This method presumes two conditions that may not always be met. The first is that the background signal has been completely removed from the component to which the compound of interest was assigned (hence, this method was not used on raw data). The second is that there are no other compounds within the subsection that have the same spectra (thereby only one chromatographic peak is assigned to a component).

Table 7.1: % RSD results for peak quantification of both the raw and IKSFA-ALS-ssel resolved data of the standards mixture injections.

	Raw Data		**IKSFA-ALS-ssel Resolved Data**		
	LCImage Software [22]	Manual Baseline	Total Sum	Manual Baseline	LCImage Software
Peak 1	4.2	3.3	4.1	1.6	9.0
Peak 2	3.8	1.8	8.4	2.2	10.5
Peak 3	13.9	12.6	19.9	4.7	34.5
Peak 4	23.2	13.1	5.8	3.5	19.4
Peak 5	12.5	5.2	6.3	1.1	1.3
Ave % RSD	**11.5**	**6.5**	**8.9**	**2.6**	**15.0**

The results for the raw data, show that the manual baseline method (average % RSD = 6.53) is almost twice as precise as those results provided by the LCImage software (average % RSD = 11.5). The IKSFA-ALS-ssel results show that the manual baseline method (average % RSD = 2.61) is four times more precise than the total sum method (average % RSD = 8.9) and greater than six times more precise than the results obtained with the LCImage software (average % RSD = 15). The poor quantitative results obtained using the LCImage software are likely due to the fact that the background correction method used in the LCImage software assumes that there are regions available with significant stretches of flat baseline that can be used to project baselines under real peaks [112]. While this assumption can be quite true of typical GC × GC data with either FID or MS detection, it is not true of LC × LC data with UV detection. The reason for this difference is due to the detector's sensitivity to refractive index changes and to the noise associated with the fast second dimension gradients. The poor results of the total sum method are due to incomplete resolution of the peak from the background signal. In other words, varying amounts of background signal are assigned to the component of the peak of interest for

each of the different sample injections and these varying intensities are summed for each of the sample injections.

Quantification of both the raw data and the corresponding chemometrically analyzed data was accomplished using the manual baseline method and LCImage software. This was done to verify that the precision of quantification was enhanced by the chemometric analysis. When the precisions for the raw data versus the chemometrically analyzed data were compared, Table 7.1, a 2.5 fold improvement in precision of quantification for the chemometrically resolved data was found when the manual baseline method was employed. The LCImage based method did not show such a dramatic improvement, probably due to the inherent differences between GC × GC and LC × LC data, as mentioned above [112].

7.3 Overview of the Data Analysis Method

Many of the issues that affect the quantification of complex samples by LC × LC analysis are not readily apparent by visual inspection of the relevant contour plots. For example, visual inspection of Figure 7.1, (urine control sample) before chemometric analysis readily leads to the inaccurate conclusion that this section consists of approximately eighteen observable compounds, some of which are well resolved, some overlapped and of varying concentrations. Upon analysis of this section by the multivariate curve resolution method used here over fifty peaks were found, and precisions (< 15 % RSD) for thirty-four peaks were determined. Of the thirty-four quantified peaks, sixteen were found only after chemometric analysis and are thus denoted by N (newly found) before the peak number. The results for these thirty-four peaks for all fourteen replicate sample injections are shown in Table 7.2. However, these results were only obtained after all of the above mentioned challenges to the data analysis (large dynamic range in

concentrations, inadequate background removal and rank deficiencies; *i.e.*, chromatographically overlapped peaks or peaks with the same spectra) were adequately understood and at least partially addressed.

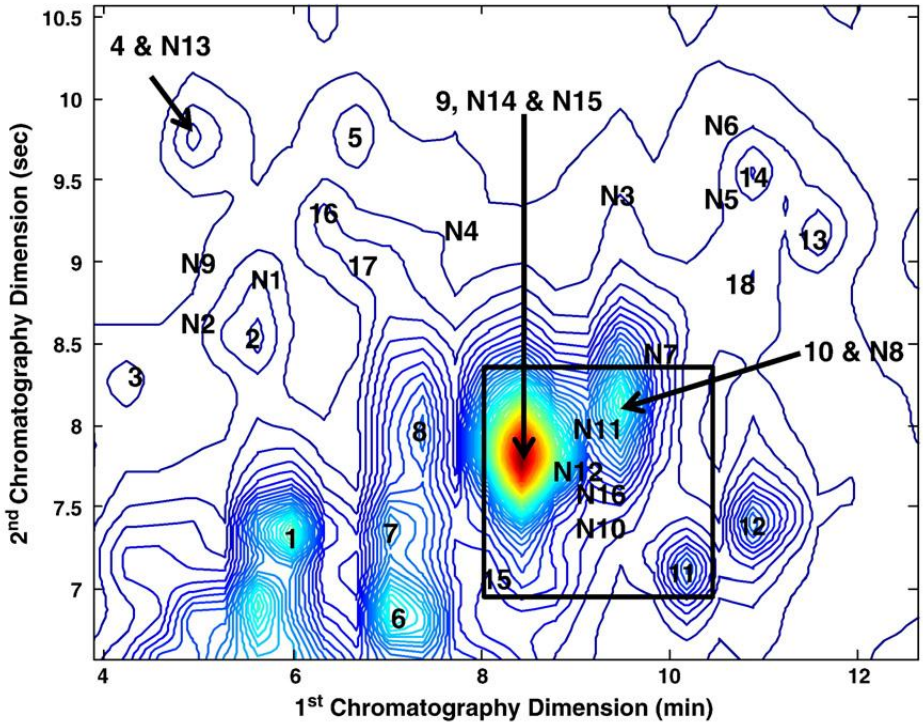

Figure 7.1: Contour plot of the urine control at 216 nm showing 34 resolved peaks. The N preceding 16 of the 34 resolved and quantified peaks signifies that those peaks were found and resolved only after application of the developed chemometric method (newly found) while the other 18 peaks were visually observable prior to chemometric analysis.

An example of a straightforward subsection analysis in which data complexities did not affect the chemometric analysis is shown in Figure 7.2. Peak 13 is the target peak of this subsection and hence it was centered within the subsection. The results of IKSFA-ALS-ssel analysis (see Figure 7.2B) show that the three major observable peaks in this subsection, peaks 13, 14 and 18 were assigned to separate components 4, 2 and 1, respectively; and the background was assigned to components 3 and 5. Peaks 13, 14 and 18 have different second dimension

Table 7.2: Precision of peak quantification of urine control sample

Visually Observed Peak Numbers	1st and 2nd Dimension Retention Times (mins)	(secs)	% RSD	Additionally Resolved Peak Numbers	1st and 2nd Dimension Retention Times (mins)	(secs)	% RSD
1	5.98	7.33	3.16	N1	5.60	8.93	2.91
2	5.60	8.58	2.73	N2	4.90	8.65	5.29
3	4.20	8.38	1.45	N3	9.45	9.53	1.56
4	4.90	9.78	1.85	N4	7.70	9.28	4.24
5	6.65	9.83	5.69	N5	10.50	9.53	4.01
6	7.00	6.93	2.98	N6	10.50	9.93	2.51
7	7.00	7.45	5.16	N7	9.80	8.53	11.10
8	7.35	7.95	15.89	N8	9.45	8.18	1.63
9	8.40	7.85	2.07	N9	4.90	9.08	0.90
10	9.45	8.18	1.96	N10	9.10	7.43	3.82
11	10.15	7.13	1.04	N11	9.10	8.05	2.25
12	10.85	7.13	3.58	N12	8.75	7.80	3.27
13	11.55	9.23	1.33	N13	4.90	9.78	1.30
14	10.85	9.58	4.16	N14	8.40	7.85	8.69
15	8.05	7.13	4.80	N15	8.40	7.85	2.23
16	6.30	9.38	3.37	N16	9.10	7.78	2.43
17	6.65	9.05	1.32				
18	10.85	9.05	3.56				

Peak number nomenclature, first and second dimension retention times for the 7th replicate sample injection and % RSD results for the 34 IKSFA-ALS-ssel resolved peaks of the urine control standard sample for all 14 sample injections.

retention times, but most importantly from the analysis point of view, different spectra, such that each peak was assigned to a different component. It is also important to note that there were no weakly absorbing peaks found as a result of the chemometric analysis of this subsection. This is the best case scenario for chemometric resolution of peak 13, in that the target peak is chemometrically resolved from the other peaks in the subsection and from the background. This leads to precise quantification using the manual baseline method. The % RSD of peak 13 was 1.33%. There are several points of interest in the analysis of this subsection. Note that two

peaks were assigned to component 1, peak 18 and a second truncated peak (the starred peak, in Figure 7.2A having a first dimension retention time greater than 12.5 min.). This truncated peak was at the edge of the selected data analysis section, and therefore not analyzed. In this work, the assignment of two peaks to the same component occurs because both peaks have the same or extremely similar spectra. However, in the case of component 1, the two peaks are chromatographically resolved thus spectral overlap is not a problem.

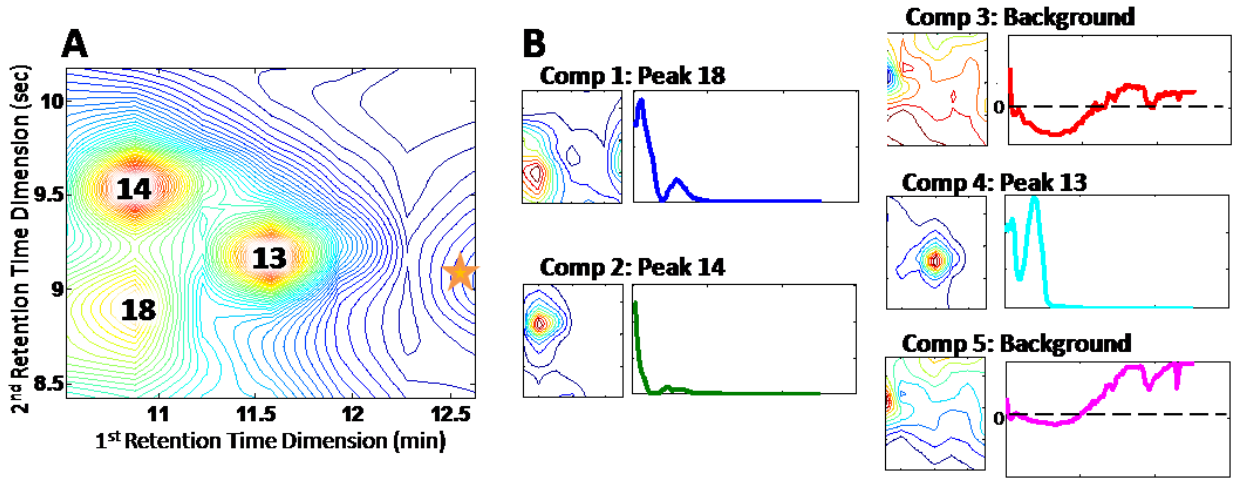

Figure 7.2: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of peak 13 subsection of urine control data. (B) Chromatographic and spectral IKSFA-ALS-ssel results for a 5 component model. This figure illustrates data that are not rank deficient in either the chromatographic or spectral dimensions. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm. The star denotes a cut off peak that was not analyzed and therefore not assigned a peak number.

7.4 Spectral and Chromatographic Rank Deficiencies (similar spectra and similar retention times)

Data are rank deficient when two or more components have the same or very similar properties in one or more data dimensions [2, 113]. Therefore, a data subsection that consists of two or more chromatographic peaks that have the same spectra is spectrally rank deficient; and a data subsection that consists of two or more chromatographic peaks that coelute with the same first and second dimension retention times and peak shapes is chromatographically rank

deficient, that is chromatographically overlapped. Data exhibiting true rank deficiency (*i.e.*, identical spectra or identical retentions) cannot be resolved using MCR methods. However, resolution of such components is sometimes possible if both forms of overlap do not occur simultaneously and there are at least some small differences in the retention or spectroscopic behaviors of the two components, *i.e.*, if two compounds with very similar, but not identical, retention times have different spectra or if two compounds with similar spectra have different retention times. Figures 7.3 and 7.4 represent cases of near spectral rank deficiency (similar

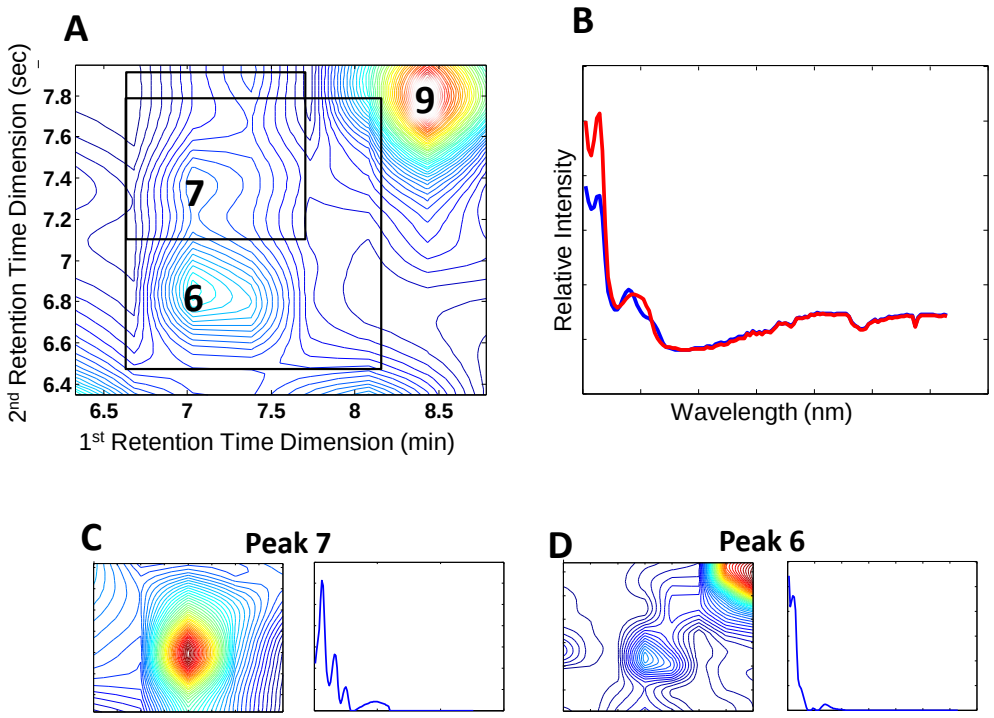

Figure 7.3: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of urine control data encompassing peaks 6 and 7 which have different 2nd dimension retention times, the same 1st dimension retention time and very similar spectra. The 2 boxes show the 2 different subsections used to separately analyze each of the peaks. (B) Overlay of the corresponding raw spectra for peak 6 (dashed line or red spectrum) and peak 7 (dotted line or blue spectrum) measured at the corresponding peak maxima, illustrating the spectral similarity of peaks 6 and 7. (C) Chromatographic and spectral IKSFA-ALS-ssel results for the component that contained peak 7. (D) Chromatographic and spectral IKSFA-ALS-ssel results for the component that contained peak 6. The chromatographic axis labels in C and D are the same as those in A, and the wavelength range is 200–700 nm.

spectra) and near chromatographic rank deficiency (chromatographically overlapped peaks), respectively. In Figure 7.3A, it is important to note that peaks 6 and 7 have the same first dimension retention time but different second dimension retention times and that peak 6 is significantly larger than peak 7. Upon chemometric analysis, peaks 6 and 7 were not resolved due to the spectral similarity of these two peaks. The raw spectra found at the apices of peaks 6 and 7 (Figure 7.3B) were shown to be very similar. Despite this, resolution of peaks 6 (Figure 7.3 D) and 7 (Figure 7.3 C) was achieved by creating two smaller subsections, one for each peak, thereby minimizing the contribution from the peak that was not of interest to the analysis.

Figure 7.4A shows a subsection that is severely chromatographically overlapped. The contour plot of the raw data reveals only peaks 9 and 10. Upon chemometric analysis (see Figure 7.4B) we now see three components (peaks 9, N14 and N15) at the first dimension retention time of 8.8 minutes. Differing IKSFA-ALS-ssel analysis models (from four to seven components) consistently revealed these three peaks. Each of these models assigned unique spectra to peaks 9, N14 and N15 (see Figure 7.4B). These three peaks are clearly not spectrally rank deficient, but they are nearly chromatographically rank deficient. Also, the second dimension peak maxima of each of the three peaks for all fourteen injections were determined. In only one of the fourteen runs did two of the three peaks show the same second dimension retention time with the greatest shift between replicate injections being 0.3 seconds. It is also noteworthy that an additional peak (N12), not observable in the raw data contour plot, is also assigned to the spectral component of peak N14. This peak was quantified using a different subsection. Thus, for this nearly chromatographically rank deficient subsection, peaks 9, N14

and N15 which have the same first and similar second dimension retention times were chemometrically resolved due to their unique spectra.

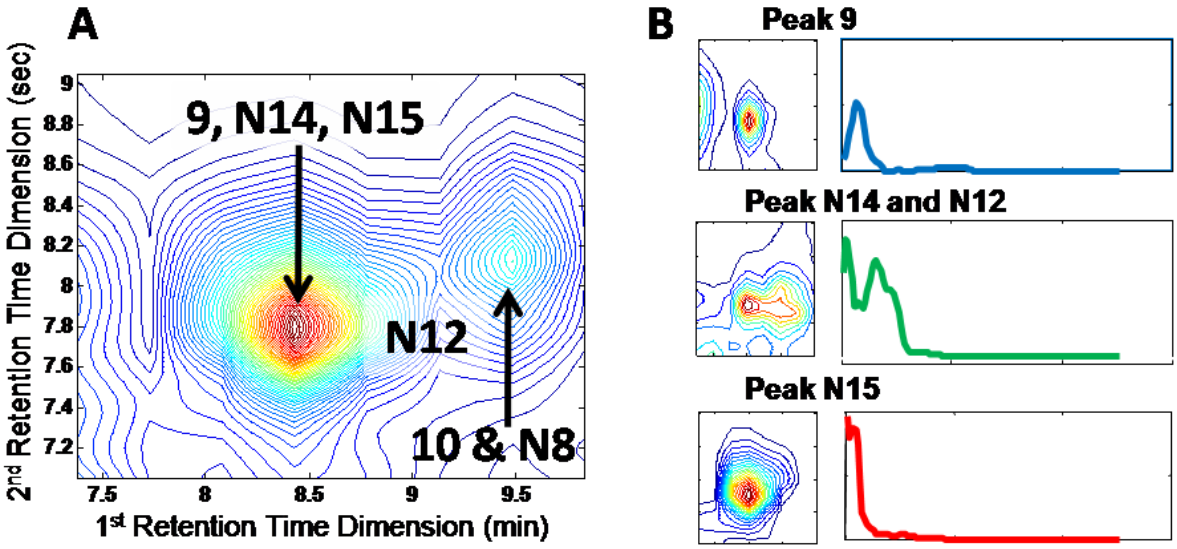

Figure 7.4: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of the subsection for peak 9. (B) Chromatographic and spectral IKSFA-ALS-ssel results showing the unique resolved spectra for peaks 9, N14 and N15 that appear to have the same first and second dimension retention times. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm.

7.5 Retention Time Shifts

Under conditions of proper sampling in LC × LC, each first dimension peak will be sampled several times and thus are present in two or more sequential second dimension chromatograms, ideally appearing at a constant second dimension retention time [19, 114]. Retention time variations, between replicate injections in the first dimension, result in changes in the sampling phase of the first dimension peak [50, 115]. The sampling phase (ϕ) as defined by Seeley [115] relates the peak maximum to the center of the sampling period in the following manner:

$$\Phi = (T\text{-}t_R)/\tau \qquad (7.1)$$

where T is the center of the sample cycle nearest to the peak maximum, t_R is the peak maximum and τ is the second dimension run time (modulation period). This concept is graphically illustrated in Figure 7.5. The red peak elutes in slices 2, 3 and 4 with the peak maximum

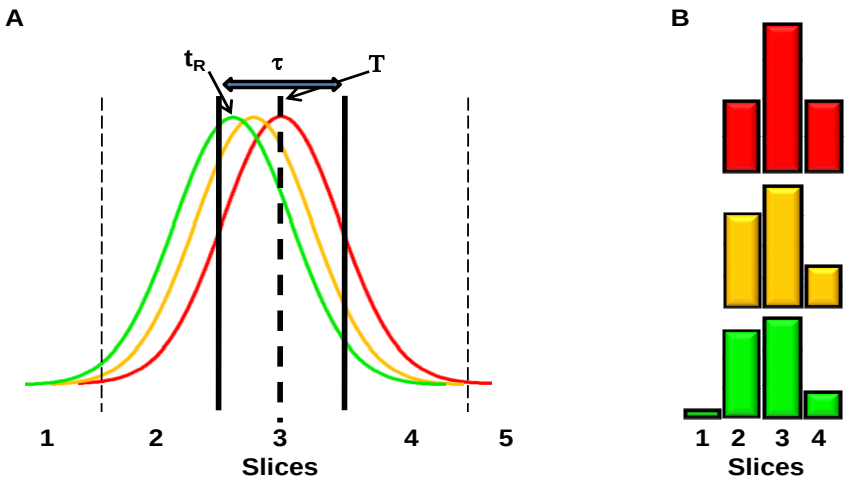

Figure 7.5: Illustration of phase shifting of ^{1}D peak. (A) The red peak simulates an exactly in phase first dimension peak having a max centered within slice 3. The yellow and green curves are peaks that have shifted earlier in the retention time but have not shifted to an exactly out of phase position. (B) Histogram representation of the area under the curve for each of the three represented peaks.

located at the center of the sampling cycle (T). Under these conditions the sampling phase is equal to zero and is said to be exactly in phase. As the sampling phase shifts, the yellow and green chromatograms, the peaks elute in a different manner and eventually even in different slices as is shown in Figure 7.5B. These differences in sampling phase between sample injections can complicate quantification.

An example of the effect of a first dimension retention time shift on partially resolved peaks is shown in Figure 7.6. Figure 7.6A gives the contour plots of the IKSFA-ALS-ssel resolved subsection containing peaks 11 and 12 for sample injections 1 and 7 in which the effect of a shift in the first dimension retention time is apparent. Figure 7.6B is an overlay of the first

(blue or dashed curve) and seventh (black or solid curve) sample injection of sequential second dimension chromatograms, slices. In this case, the peak of interest, peak 12, is not resolved from peak 11 because these peaks are spectrally rank deficient. The peak in the first slice of the first sample injection chromatogram in Figure 7.6B corresponds to peak 11; while peak 12 is seen in slices 3, 4 and 5. In this sample injection, peaks 11 and 12 were separated by the fortuitous timing of the valve switching at the minimum between the two peaks. However, in the seventh sample injection chromatogram, peaks 11 and 12 coelute in slice 2 due to the changes in the first dimension retention time. In this case, phase shifting causes the first slice of peak 12 to coelute with the last slice of peak 11. This coelution leads to irreproducibility in the area determination due to inadequate resolution of the two peaks in the second dimension. A schematic representation of this issue is presented in Figure 7.6C, depicting two sample injections phase shifted relative to one another (sample injection 1 and sample injection 7, respectively). Time points where the valve is switched are shown by the vertical lines. In injection 1, the valve switches position at the minimum between the two peaks; while in injection 7, the positioning of the valve switch results in a second dimension chromatogram that encompasses the tailing end of peak 11 and the leading edge of peak 12. The resolved and unresolved bar graphs in Figure 7.6C can be thought of as corresponding to injection 1 and injection 7 of Figure 7.6A and B, respectively. If the first slice of injection 7 is compared with the bar graph, it is clear that the co-eluting peak contains area belonging to both peak 11 and peak 12 making a clean integration of peaks 11 and 12 impossible due to retention time shifting over the course of a long series of sample injections. Due to the use of a spectral curve resolution approach this is only a problem for spectrally similar (*i.e.*, spectrally rank deficient) peaks; clearly without the use of a curve resolution approach such as IKSFA-ALS-ssel, the problem would be even more severe.

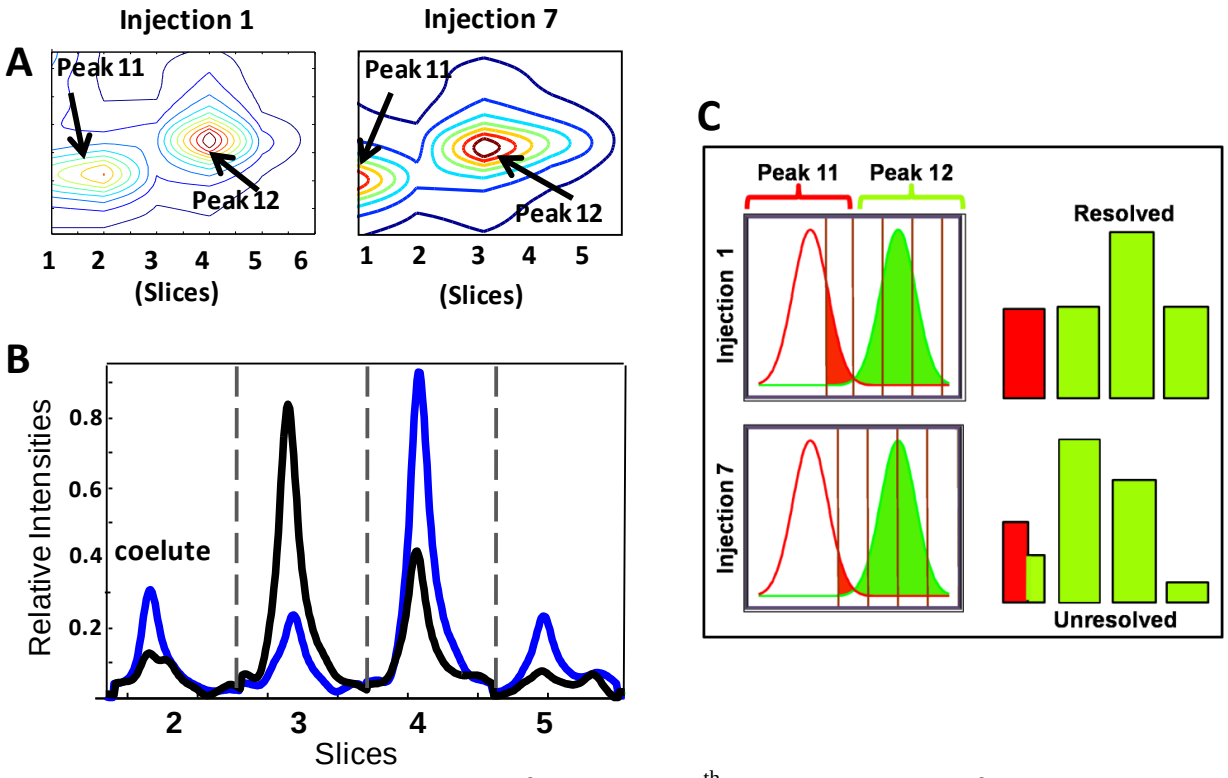

Figure 7.6: (A) Contour plots at 216 nm of the 1st and 7th sample injections after multivariate analysis of the peak 12 subsection. (B) Overlay of the sequence of 2nd dimension chromatograms after IKSFA-ALS-ssel analysis such that the blue or dashed line chromatogram corresponds to sample injection 1 and the black or solid line chromatogram corresponds to sample injection 7 which shows the coelution of peaks 11 and 12 owing to phase shifting in sample injection 7. (C) Schematic representation of the effects of phase shifting on the quantitative analysis of chromatographically overlapped peaks.

7.6 Dynamic Range Issues

Another complicating issue is directly related to the large dynamic range in compound concentrations of the samples typically encountered with LC × LC analysis of metabolomics samples. This is the case for the analysis of peak N16 (see Figure 7.7, such that N denotes a peak found only after chemometric resolution and not visually observed in the contour plot) which is assigned to component 1 in Figure 7.7B. A set of contour plots of the fourteen replicate injections for component 1 is shown in Figure 7.8A; and at least two additional smaller peaks to either side of peak N16 can be seen. The intensities of these two weaker peaks are

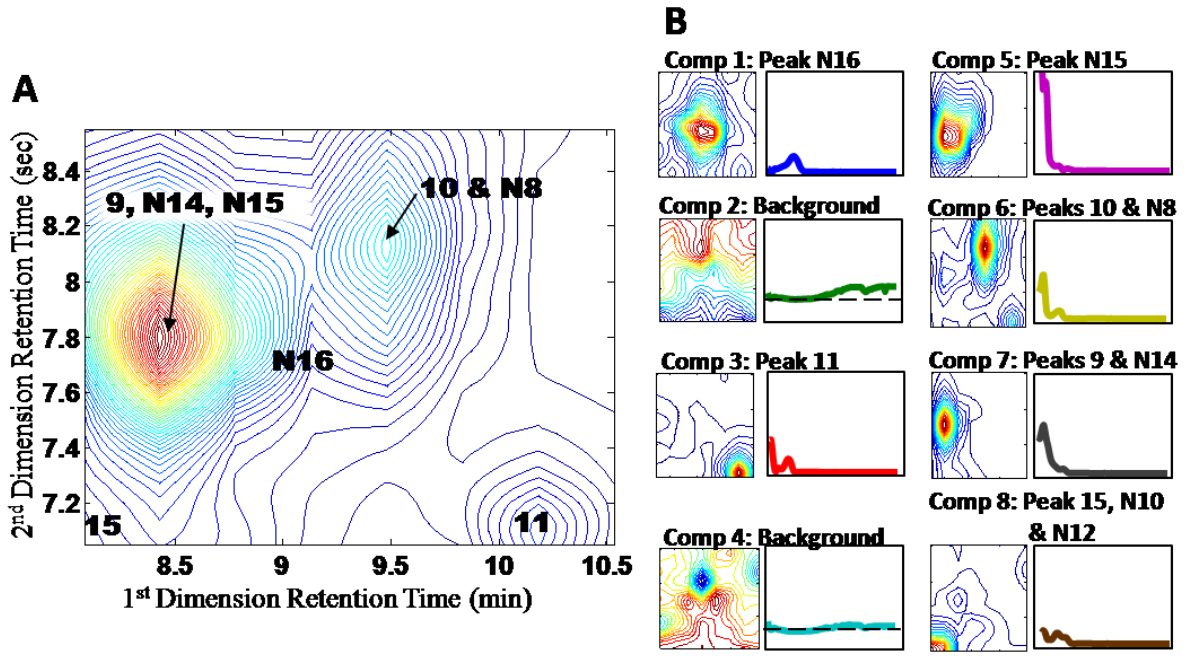

Figure 7.7: (A) Contour plot at 216 nm of the 7th sample injection before multivariate analysis of peak N16 subsection of the urine control data. (B) Chromatographic and spectral IKSFA-ALS-ssel results for an 8 component model. The chromatographic axis labels in B are the same as those in A, and the wavelength range is 200–700 nm.

approximately five times less than that of peak N16, which is already substantially lower in intensity compared to nearby peaks 9, N14 and N15. These two additional peaks embedded under peak N16 and to either side of it have either the same or very similar spectra as Peak N16. Specifically, one unresolved, low concentration peak begins eluting in the first dimension approximately one slice (second dimension run) before peak N16 and at approximately the same second dimension retention time as that of peak N16; and a second unresolved, low concentration peak begins eluting in the first dimension approximately one slice after peak N16 and at approximately the same second dimension time. These embedded peaks are not always obvious by inspection of the raw chromatograms of the fourteen replicate injections. From observations of injections 6-8, it is difficult to determine if the asymmetry of the peak was due to the effects of retention time shifts or due to the presence of embedded peaks.

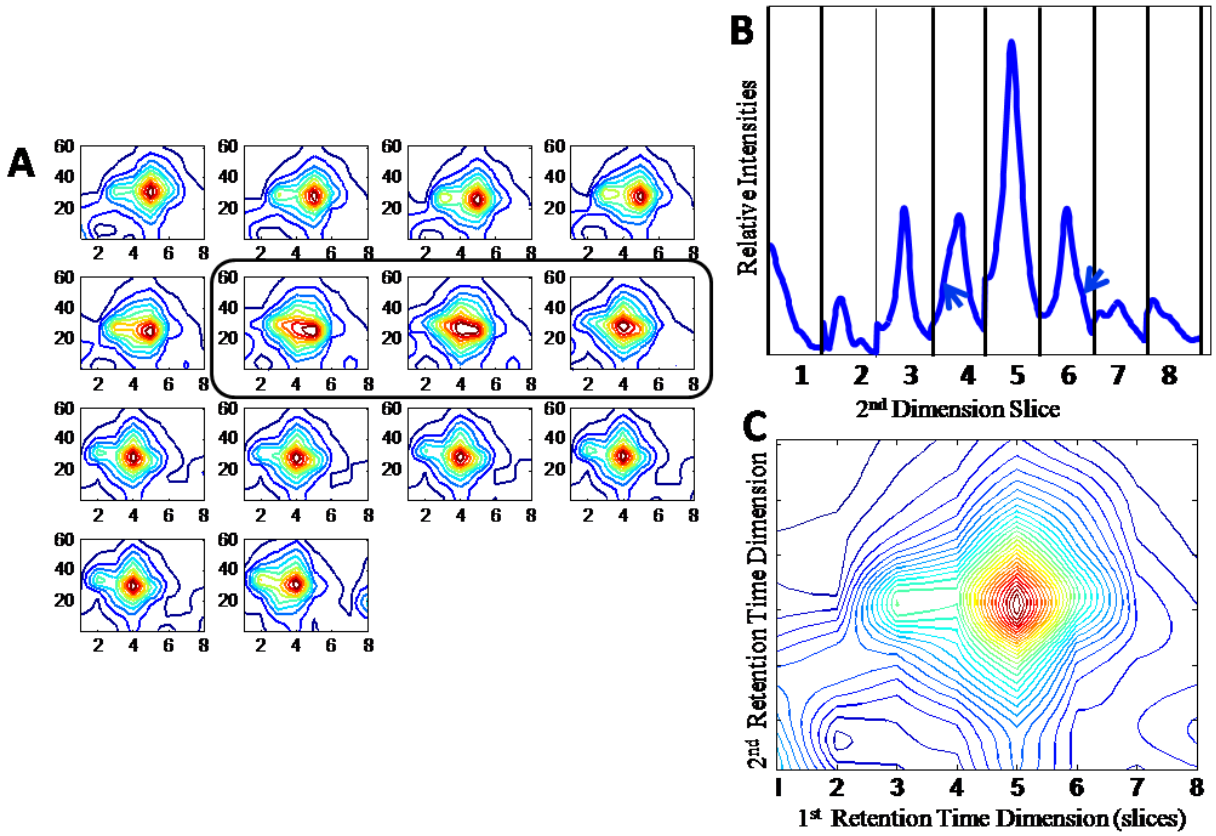

Figure 7.8: (A) Contour plots at 216 nm of the IKSFA-ALS-ssel resolved component for the analysis of peak N16 for all 14 sample injections where injection 1 is the top left hand corner and injections follow sequentially to injection 14 in the bottom right hand corner. Injections 6–8 are within the rectangular box. (B) Sequential 2nd dimension chromatograms for sample injection clearly indicating the presence of "embedded" peaks in the 4th and 6th slices as shown by the arrows. There are 61 data points for each 2nd dimension slice and 8 1st dimension data points for a total of 489 data points on the sequenced chromatograms. (C) Corresponding contour plot at 216 nm for injection 1. The chromatographic axis labels in A are the same as those in C, and the wavelength range is 200–700 nm.

Another way to find embedded peaks is to examine the sequence of second dimension chromatograms (see Figure 7.8B), note that the arrows indicate deviations from a Gaussian peak shape. By comparing the sequential chromatogram with the contour plot of the first sample injection (Figure 7.8C), a truncated peak in both chromatographic directions is observed in the first slice of the subsection. The second slice indicates that there is a very small component eluting quite early in the second dimension and then the obvious first embedded peak is seen. The third slice has a chromatographic peak belonging to the first embedded peak while the fourth

slice shows a chromatographic peak exhibiting a shoulder on the left of peak N16 corresponding to the first embedded peak. The fifth slice consists of peak N16 at its maximum while the sixth slice now shows a chromatographic peak exhibiting a shoulder on the right of the peak corresponding to the second embedded peak mentioned above. From the chromatogram in the Figure 7.8B, it also became apparent that there is yet another peak eluting very early in the seventh and eighth slices; this is seen more clearly in the contour plot of the fourteenth injection, Figure 7.8A. So for just component 1 of the eight component model for the analysis of peak N16, we have shown that there are actually five incompletely resolved compounds. This lack of resolution is directly related to the relative intensity differences of the peaks. The large dynamic range issue associated with the two embedded peaks and peak N16, makes resolution of these three peaks with current chemometric methods very difficult. From a chromatographic standpoint, a more selective detector, such as a mass spectrometer, may aid or completely alleviate the issue of dynamic range.

7.7 Inadequate Background Removal

Preliminary IKSFA-ALS analysis sometimes shows a background component that has a negative peak at the same retention time as the compound of interest. From Figure 7.9A, where peak 5 is used as an example of this effect, it is clear that the relative intensity of the compound of interest will be adversely affected by the negative peak in the background component because of the lack of resolution of the peak from the background. We found that this problem can often be ameliorated by using the spectral selectivity (ssel) constraint of the ALS algorithm as was previously described [1]. Briefly, the spectra of all non-background components are constrained to be zero from 440 to 700 nm, and the non-negativity constraint is applied to all spectral components which correspond to real chemical species. These constraints were imposed because

background spectra dropped below zero and differed significantly from the analyte spectra at wavelengths greater than 440 nm. This is principally a result of refractive index changes associated with gradient elution. In contrast, none of the real chemical constituents that we observed to be present in urine absorb at wavelengths greater than 440 nm. Therefore, these constraints allowed the algorithm to more accurately resolve background components from real peaks. The results of the implementation of the spectral selectivity constraints in this manner are shown in Figure 7.9B.

7.8 Additional Issues

To further illustrate the complexity of the data, with respect to the number of chromatographic peaks found to be present, the small subsection chosen to analyze peak N16, Figure 7.7A, will be discussed. Inspection of the single wavelength contour plot of the raw signal, suggests that the region around N16 is relatively uncomplicated with only three peaks; however, upon application of IKSFA-ALS-ssel, which indicates that eight components exist (although two are assigned as background components), it became apparent that there are at least six peaks within this small subsection. Peak N16 is not observable in the raw signal contour plots, but it is detected upon curve resolution. The spectral and chromatographic results for the analysis of this subsection are shown in Figure 7.7B. A detailed analysis of peak N16 (component 1) described in section 7.6, showed that this component consists of five unresolved compounds.

Table 7.3 shows the peak count after this same procedure was followed for additional subsections for each of the compounds found within the original peak N16 subsection. It was determined that component 3 in Figure 7.7B has two compounds (peak 11 and an embedded peak; *i.e.*, a very small peak with a similar retention time as a larger peak). Component 5

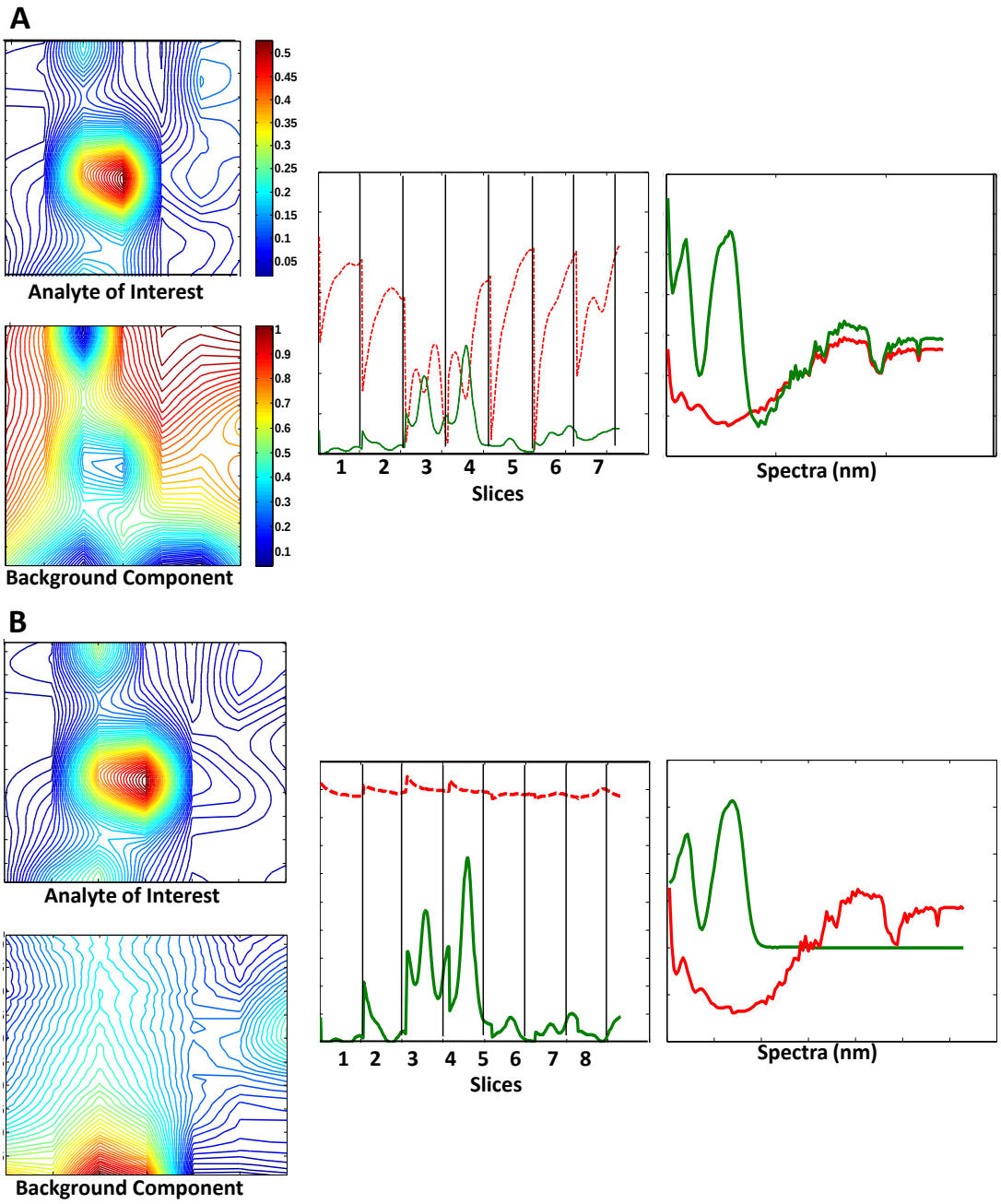

Figure 7.9: Peak splitting that results in a negative peak that corresponds to the analyte of interest in the background component. (A) Contour plots of the peak of interest and of the background after multivariate analysis without implementation of the spectral constraints, along with corresponding overlay plots of the sequence of 2nd dimension chromatograms and spectra. (B) Contour plots of the peak of interest and of the background after IKSFA-ALS-ssel, along with corresponding overlay plots of the sequence of 2nd dimension chromatograms and spectra. The dashed curve corresponds to a background component, and the solid curve corresponds to the component of interest.

contains only peak N15. This peak is resolved in component 5 for this analysis but is the overlapped peak associated with the resolution of peaks 9, N14 and N15 in section 7.4. There are two observable peaks present in component 6 and following a separate analysis of peak 10, an additional peak at the same first dimension retention time as peak 10 was also found (peak N8), along with three embedded peaks. Therefore, component 6 consists of a total of five compounds. Component 7 appears to consist of only peak 9. From the chemometric analysis of a subsection specifically for peak 9, see section 7.4, we determined that there are three compounds at this first dimension retention time with three different spectra such that peak 9, peak N14 and peak N15 are chromatographically severely overlapped. In the fit results shown in Figure 7.7, peaks 9 and N14 share component 7 in that these peaks are not resolved with respect to each other, and peak N15 was assigned to component 5. There are three distinguishable compounds present in component 8. The major peak observed in component 8 is

Table 7.3: Combined analysis results of several smaller subsections showing all of the detected peaks that were found in the subsection used for the analysis of peak N16.

Spectral Component #	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Resolved Peak #	N16	bkgd[a]	11	bkgd[a]	N15	10 N8	9 N14	15 N12 N10
Embedded peaks	4		1		0	3	0	0
Total # of peaks found	5		2		1	5	2	3

[a]Background contribution

peak 15. The two minor peaks were resolved and are labeled as peaks N10 and N12. Therefore, a total peak count for this small subsection, 2.8 minutes by 1.5 seconds (first retention time dimension by the second retention time dimension), meant to analyze peak N16, was determined

to consist of 18 compounds. The effective peak capacity for this subsection was determined to be 5.3 using the D-S-C model as described by Davis *et al*. [116]. However, with the use of chemometric tools, we are able to quantify 10 peaks, and to detect 8 additional peaks, which is well beyond the peak count considered feasible using peak capacity as a metric.

7.9 Conclusions

The ability to precisely and easily quantify LC × LC data will enable this important analytical technique to be applied to the analysis of increasingly more complex samples that are of interest in the various –"omics" fields. In this work we have explored various key issues, addressed some of them through the application of the IKSFA-ALS method and have made specific recommendations for how the remaining issues might be approached in the future.

Spectral and chromatographic rank deficiencies: As reported here, near rank deficiencies in one of the dimensions can be addressed using the IKSFA-ALS algorithm. Additionally, other work in this laboratory addresses spectral rank deficiencies using a novel unimodality constraint [105]. Mass spectral detection, as opposed to DAD, is likely to decrease the incidents of spectral rank deficiency, albeit at a cost of the precision of quantification. However, severe chromatographic rank deficiencies in both chromatographic dimensions must be addressed by either improving the chromatographic selectivity or by sample pretreatment to simplify the matrix.

Retention time shifts: One of the advantages of the IKSFA-ALS method as compared to PARAFAC [10] is that retention time shifts do not cause severe problems for compounds that have unique spectral characteristics. However, for compounds that have both exceptionally similar spectral and chromatographic characteristics, we have shown that shifts in the first dimension retention time can cause significant variations in resolution as the relative sampling

phase changes from sample to sample. The issues related to retention time shifting may be better understood by use of computer simulations and not by real data. Experimental as opposed to computational strategies, such as improved temperature and flow control, can increase retention time reproducibility to reduce this type of difficulty. In addition, once retention times become more reproducible, chemometric methods, such as PARAFAC [10] that are more easily automated than IKSFA-ALS, can then be employed. Clearly, improving retention reproducibility remains a high priority for optimizing the quantitative aspects of LC × LC.

Dynamic range issues: Analysis of complex samples with many peaks over a large dynamic range of concentrations remains a problem. Here, we recommend enhancing peak capacities as well as improving chemometric methods *per se* to address the analysis of such samples. In lieu of these advances, additional sample pretreatment steps should be developed to satisfactorily address these issues; *e.g.*, in proteomic studies, the most common high abundance proteins are removed prior to analysis [117, 118].

Inadequate background removal: Both the reproducibility and magnitude of the background contributions influence how well the IKSFA-ALS algorithm can resolve the background from the analyte signals. Development of detectors with less sensitivity to refractive index changes [119, 120], as well as development of instrumental improvements that lead to better temperature and flow stability are recommended. In addition, strategies for background removal that do not include curve resolution may also be useful [22].

Many of the issues affecting quantification addressed in this work will be most effectively addressed by chromatographers and chemometricians working together (1) to improve the instrumentation, to make better use of the available separation space and to provide improved long term reproducibility of retention, and (2) to improve the data analysis, to develop

algorithms that are not restricted to multilinear data, that better handle rank deficient data and that are more user-friendly.

Chapter 8: Comparison of Chemometric Methods for the Screening of Comprehensive Two-Dimensional Liquid Chromatographic Analysis of Wine

This chapter investigates the use of two different chemometric methods, the Fisher ratio (FR) method and the similarity index (SI) method (supervised and unsupervised, respectively) for the rapid screening of comprehensive two-dimensional liquid chromatographic (LC × LC) analysis of wine. To the authors' knowledge, neither of these methods has been used in the analysis of LC × LC data in which diode array detection was employed. An experimental data set consisting of five different wine samples and a simulated data set were analyzed in the investigation of these screening methods. The previously developed IKSFA-ALS-ssel method was used to resolve and to quantify three peaks giving a most dissimilar SI result, three peaks with a most similar SI result and three peaks that appeared to lie somewhere in between the most and least similar set of peaks, for a total of nine peaks. The determined relative concentrations of these nine peaks were used in the validation of the screening methods. To further the understanding and verification of the results of the similarity index and Fisher ratio methods, the following statistical analyses were employed: the Tukey-Kramer honestly significant difference (HSD) test, an equivalence test and ANOVA. The goal was to determine the applicability of the chemometric methods for the analysis of complex four-way data for the rapid screening of multiple wine samples to locate the peaks that represent significant concentration differences

between the samples. These methods provide two significant advantages. The first is that an ever changing model based on vintage or any other parameter is not required, as was discussed in Chapter 4.2. And the second is the elimination of the time consuming requirement of identification and quantification of all compounds in every sample being analyzed, by only requiring the further analysis of the minimal number of compounds found to be significantly different.

8.1 Theory

8.1.1 Alignment algorithm

In order to align the second dimension retention times between sample injections, a global shift parameter for adjusting each sample chromatogram is determined. The first step is to determine the position of the maximum in the second retention time dimension for several strongly absorbing peaks that appear in all sample injections. The sample injection with the earliest eluting second dimension retention time is used as a reference point for all of the peaks and the change in retention time with respect to the reference is determined for all of the peaks for all sample injections. The average change in retention time for all peaks is calculated for each sample injection. This value is the per injection global shift parameter for each of the sample chromatograms. The maximum and minimum value of the shift parameter across all samples is determined. Each sample chromatogram is then essentially shifted in the second retention time dimension by removing the same total number of data points from the beginning and/or the end of each sample injection using the shift parameter, as illustrated in Figure 8.1. The starting point in the second dimension for each sample injection is determined by subtracting the minimum value of the shift parameter from each index parameter value plus one. The end

point for each sample injection is determined by subtracting the maximum value of the index parameter from the index parameter plus the total number of second dimension data points.

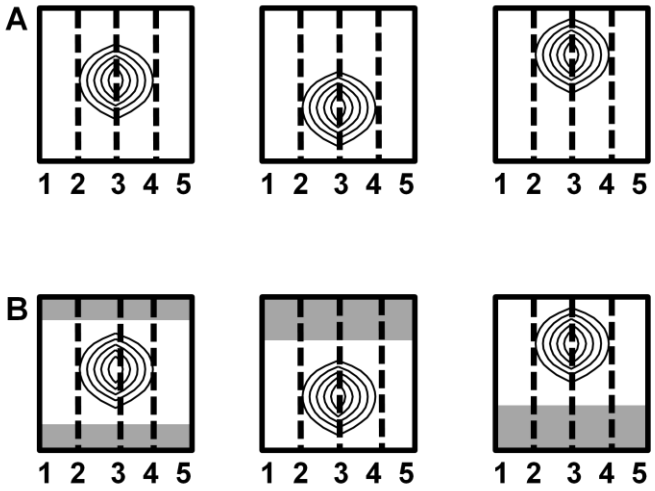

Figure 8.1: Schematic representation of global alignment applied in the 2nd retention time dimension. (A) Three sample injections illustrated with only retention time shifting in the second dimension. (B) Each sample injection is aligned using a global parameter such that the maximum for each peak is in the same position and the data size dimensionality remains consistent.

8.1.2 Similarity Index Method (SI)

Windig [121] described several methods for reducing liquid chromatography mass spectrometry (LC/MS) data to only those mass chromatograms that show differences among them, thus reducing the total analysis time required for such complex data. This was accomplished using a ranking system, termed the similarity index, from zero to one. Here we describe the COMPARELCMS_SIM algorithm, which will be referred to throughout as the SI (Similarity Index) method, as implemented not for LC/MS data but for LC × LC-DAD analysis, where the data array $\underline{\mathbf{X}}$ has dimensions $I \times J \times K \times L$. Here, I is the number of data points in each second dimension chromatogram, J is the number of data points in each first dimension chromatogram, K is the number of different samples that were analyzed and L is the number of points in each spectrum. The data are unfolded, combining the two chromatographic dimensions

such that $P=IJ$, and arranged such that $\underline{\mathbf{X}}$ is now $K\times L\times P$. The mean value (equation 8.1) and the minimum value (equation 8.2) of the K different samples of data array $\underline{\mathbf{X}}$ are found as follows:

$$\mathbf{X}^{\text{mean}} = \sum_{k=1}^{K} \frac{x_{klp}}{K} \tag{8.1}$$

$$\mathbf{X}^{\text{min}} = \min_{k}(x_{1lp}, x_{2lp}...x_{klp}) \tag{8.2}$$

The correlation coefficient, r_p, is then calculated between the columns of $\mathbf{X}^{\text{mean}}$ and $\mathbf{X}^{\text{min}}$

$$r_p = \frac{\sum_{l=1}^{L}\left[\left(\frac{x_{lp}^{\text{mean}} - \text{ave}(\mathbf{X}^{\text{mean}})}{\text{std}(\mathbf{X}^{\text{mean}})}\right)\left(\frac{x_{lp}^{\text{min}} - \text{ave}(\mathbf{X}^{\text{min}})}{\text{std}(\mathbf{X}^{\text{min}})}\right)\right]}{L} \tag{8.3}$$

where r_p gives the correlation coefficient for each time point in the chromatographic data set and ave and std indicate the mean and the standard deviation of the rows of $\mathbf{X}^{\text{min}}$ and $\mathbf{X}^{\text{mean}}$. The similarity index, s_p, is then calculated by weighting the correlation coefficient by the ratio of the lengths of the $\mathbf{X}^{\text{mean}}$ and $\mathbf{X}^{\text{min}}$ vectors. This ratio corrects for the intensity differences between the minimum and mean chromatograms.

$$s_p = r_p\left(\sqrt{\sum_{l=1}^{L}(x_{lp}{}^{\text{min}})^2} \Big/ \sqrt{\sum_{l=1}^{L}(x_{lp}{}^{\text{mean}})^2}\right) \tag{8.4}$$

Peaks having a similarity index (SI) value of 1 are exactly the same in both profile and intensity, while a peak with a SI value of 0 indicates the presence of the peak in one sample but its absence in another.

8.1.3 Fisher Ratio (FR)

Typically, the Fisher ratio (FR) has been applied to two-way data and is a class based statistical analysis. Pierce *et al.* [122] applied this approach to third-order GC × GC-TOFMS data, stating that the FR is capable of distinguishing areas of the data that exhibit significant between class variations (chemical differences) relative to within class variations (random noise). The Fisher ratio is equal to the class-to-class variance divided by the within class variance such that

$$F = \sigma_{cl}^2 / \sigma_{err}^2 \tag{8.5}$$

The class-to-class variability (σ_{cl}^2) is defined as

$$\sigma_{cl}^2 = SS_{fact} / (Q\text{-}1) \tag{8.6}$$

where P is the number of classes and SS_{fact} is the sum of squares between classes

$$SS_{fact} = \sum_{q=1}^{Q} N_q \left(\bar{x}_q - \bar{x} \right)^2 \tag{8.7}$$

N_q the number of replicates in class q, $\bar{x}_q$ is the mean of the q^{th} class and $\bar{x}$ is the overall mean of the data. The within class variance squared (σ_{err}^2) is defined as

$$\sigma_{err}^2 = SS_R / (K\text{-}Q) \tag{8.8}$$

where K is the total number of sample injections and SS_R is the residual sum of squares within classes

$$SS_R = \sum_{q=1}^{Q} \sum_{n=1}^{N_q} \left(x_{nq} - \bar{x}_q \right)^2 \tag{8.9}$$

where x_{nq} is the n^{th} measurement in the q^{th} class. In order to implement the FR method, the data is reshaped as described for the SI method in the above section.

8.2 Experimental:

ACDLABS ChromProcessor 9.0 (Advanced Chemistry Development, Inc. Toronto, Canada) was used to import the experimental data into the Matlab 2007a (Mathworks, Inc. Natick, MA) environment. JMP 8 (SAS Institute, Inc. Cary, NC) was used for the statistical equivalence testing and for the Tukey analysis. LCImage software v 2.1 (GC Image, LLC Lincoln, NE) was used for peak counting [22]. All data analysis was performed on a HP Pavilion dv9500 with 4.0 GB RAM, an Intel®Core™ 2 Duo CPU T7500 at 2.2 GHz processor operating with the Windows Vista Home Premium operating system.

8.2.1 *Data organization*

Due to limitations with respect to memory in Matlab, all data analyses were performed on a subsection from 3.85 to 10.45 seconds in the second dimension (276 data points) and from 3.5 to 14.35 minutes in the first dimension (32 data points) as shown by the large box in Figure 8.2 (complete details in regard to this data can be found in Chapter 4.2). This section encompasses most of the peaks to the left of the ridge (possibly due to on-column reactions), and it does not contain a large area of empty separation space. Geographical variability was studied using the five different sample types, HRC A-C (control samples from three different batches acquired from the Horticultural Resource Center of the University of Minnesota), CV (sample acquired from the Ceplecha vineyard) and WV (sample acquired from the Winter vineyard) as discussed in Chapter 4.2; The data matrix size was 276×32×126×15 before global alignment of the data, 265×32×126×15 after global alignment and 15×126×8480 after combination of the two

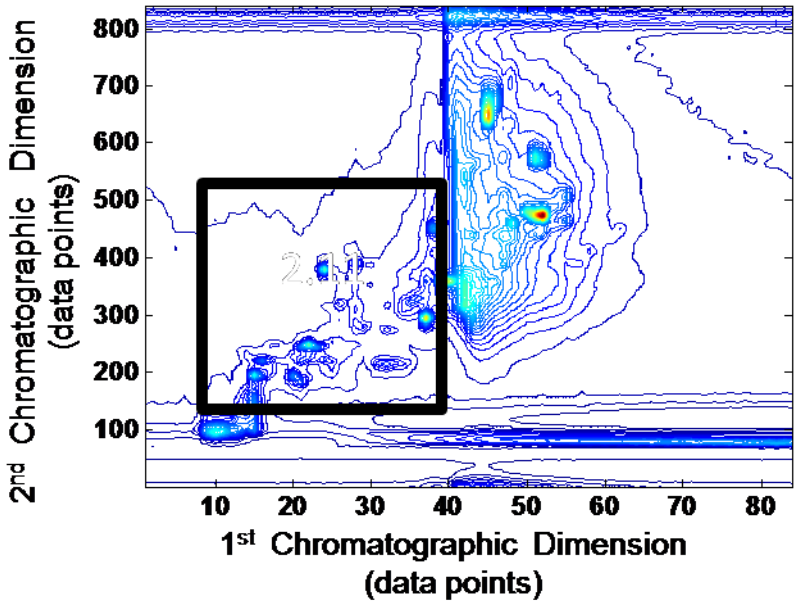

Figure 8.2: Contour plot of HRC C 1st replicate at 216 nm. The boxed area is the section of the chromatograms analyzed in this work.

chromatographic dimensions and rearrangement of the data for implementation of both the SI and FR algorithms. The resulting SI value matrix (consisting of the SI values calculated according to eq. (9.4)) and FR value matrix (consisting of the *F* values calculated in eq. (9.5)) can be plotted as two dimensional chromatographic contour plots, as shown in Figure 8.3.

The simulated data consist of nine chromatographically well resolved peaks without retention time shifts in either chromatographic dimension. The concentrations of these nine peaks were taken from the relative concentrations of the nine IKSFA-ALS-ssel quantified peaks of the experimental data. Normalized, known spectra from a standards amphetamine analysis were assigned to the nine peaks. The fifteen experimental wine samples were run with two replicates of a known standards mixture containing seven indoles (Chapter 4.2). A section of these two chromatograms was used to add a background in varying ratios to the simulated data.

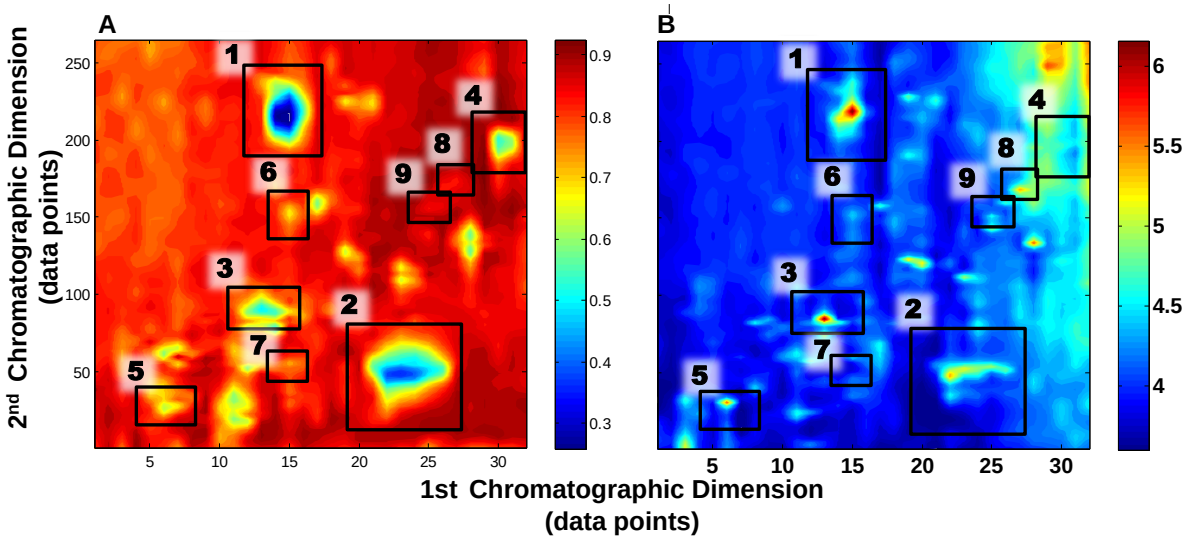

Figure 8.3: (A) Similarity Index contour plot showing each of the nine analyzed peaks and the associated scale from 0 (most dissimilar) to 1 (most similar). (B) Fisher ratio contour plotted using a logarithmic base 10 for scaling showing the nine analyzed peaks and the scale.

8.2.2 *Data analysis scheme*

A visual inspection of the 2D chromatograms indicates that there are approximately thirty peaks present. To further confirm the number of peaks in the section of data, LC Image software was used for peak detection and the peak count was found to be between 29 and 33 depending on the sample injection analyzed. This variability with regard to peak count from sample to sample is not due to the presence or lack thereof of a given compound in different samples. It was determined rather that the means by which LCImage determine peak boundaries (blobs-which is to be representative of a peak) varied between samples. For example, in a given sample injection, it was visually apparent that the LCImage software had included two different peaks within one peak boundary, thereby reducing the peak count for that sample injection.

In this work, the second dimension retention time shifted between sample injections significantly; while the first retention time dimension exhibited minimal retention time shifting. Windig [121] employed an optional two-dimensional finite impulse response (FIR) filter which

essentially smoothes the chromatograms to handle minor retention time shifting. However, this was implemented for 1D-LC data and not for LC × LC data being analyzed in this work, such that the two chromatography dimensions have been unfolded for both the SI and FR analysis purposes. Pierce *et al.* [122] state that retention time shifting affects the performance of their developed Fisher ratio method. Since both methods required prealignment, we choose to globally align the peaks in the analyzed section in the second retention time dimension prior to implementing either algorithm in an attempt to minimize dissimilarity contributions due to retention time shifting.

Geographical variability was investigated using both the SI and the FR algorithms. Threshold values for each method are needed to determine an appropriate peak exclusion/inclusion threshold for further quantitative analysis. Several factors were considered in the determination of both the similarity index threshold (SIT) and the Fisher ratio threshold (FRT). First, it was necessary to eliminate within class variability so that only between class variability was studied. Second, an attempt was made to minimize both type 1 errors (false positives, including peaks that are not significantly different) and type 2 errors (false negatives, excluding peaks that are indeed significantly different). The within class variability for each of the five classes was determined in a slightly different manner for the SI method and the FR method and is thoroughly discussed in Section 8.3.2. In an effort to minimize type 1 and type 2 errors, the most dissimilar index values for each of the five within class analyses were averaged. This average value was then used as a threshold for inclusion/exclusion from further analysis.

A range of different peaks having similarity indexes both below and above the SIT were chemometrically resolved using a previously described method, IKSFA-ALS-ssel [1]. The resolved peaks were then quantified using a manual baseline method described in Chapters 5 and

7 [13]. Using these concentrations, a simulated data set was created to investigate chromatographic conditions that may affect the results of either of the screening methods. Several statistical analyses were also performed using the quantification results. The percent relative standard deviation, % RSD, for each peak was found by dividing the standard deviation of the relative concentration of the different sample injections by the average determined relative concentration for all sample injections and multiplying by 100. The inner quartile range (IQR) and the range for each of the nine peaks were also calculated. Tukey's Honestly Significant Difference (HSD) test [123] and an equivalence test [124] were performed in an attempt to determine which peaks were statistically different or equivalent between classes. ANOVA analyses were also performed using both a three group (classifying the nine HRC control batch injections together) and a five group analysis. These results are further discussed in Section 8.3.1.

8.3 Results and Discussion:

For the sake of clarity, we have chosen to use the ranking from the SI results as a means of labeling the nine IKSFA-ALS-ssel analyzed peaks. Figure 8.3A is the SI contour plot from the analysis of all fifteen injections and shows the nine peaks selected for further study. The intensity bar associated with the SI contour plot ranges from 0 to 1 where those peaks contoured with blue (the darkest) have SI values closest to zero and those with red (the lightest) have SI values closest to one. Note that, as expected, the background regions, being very similar for all sample injections, show high SI values. The peaks are labeled such that peak 1 exhibits the most dissimilarity (with a SI value closest to zero) and peak 9 exhibits the most similarity (with a SI value closest to 1).

8.3.1 Statistical analysis of concentration data

Since the algorithms are intended to determine concentration differences between samples, nine peaks (three with the lowest SI values, three with the highest SI values and three that appeared to have SI values in between) were chemometrically resolved by IKFSA-ALS-ssel and quantified by a manual baseline method [13]. The relative concentrations for these nine peaks in all fifteen injections were determined. Using calculated concentrations as true values (which are unaffected by chromatographic influences such as background, retention time shifting and peak overlap), we sought to determine a possible metric for the statistical determination of the dissimilarity ranking of the nine peaks. To be user friendly, this metric must meet two criteria. It must be rapid, efficient and selective; and not require individual peak selection and quantification, which is fairly tedious and time consuming. The percent relative standard deviation (% RSD) was calculated for each of the nine peaks. Since the % RSD is a measure of the variation around the mean concentration, it follows that peaks with greater % RSD's would have values that are farthest from the mean and therefore the most different concentrations between samples. However, % RSD depends on the magnitude of the average concentration, as well as the standard deviation from that mean. Hence, two peaks that have a similar standard deviation but with very different average concentrations will have very different % RSD results. Specifically, two peaks may have a similar standard deviation; the peak with a high average concentration will have a lower % RSD as compared to a peak with a low average concentration that will result in a higher % RSD. This is evidenced in Table 8.1 for peaks 3 and 4 with relatively similar standard deviations and for peaks 7 and 8. Peak 4 has a lower average concentration but the corresponding % RSD is almost twice that of peak 3; hence, the

dissimilarity rank order is affected. The average concentration of peak 7 is almost four times greater than that of peak 8 and therefore has a % RSD almost four times smaller.

The total range of the concentrations for each peak and the interquartile range (IQR) were also calculated and are compared in Table 8.1. The IQR, a statistical measure of variability, is more robust than the total range because it uses the middle 50% of the data and is therefore not affected by outliers. However, in the case of all fifteen sample injections, nine out of the fifteen should be (and in fact are) relatively similar because these samples are from the same batch of wine, *i.e.,* the HRC A-C control samples. The IQR will also not represent the data range adequately if the Ceplecha vineyard and Winter Vineyard samples have concentrations above and below the HRC control samples since the first quartile will eliminate the low concentrations and the third quartile will eliminate the high concentrations leaving the three similar within batch samples. Since it is reasonable from reviewing the concentration data to assume that there are no outliers, the total range of the data is the better metric in the determination of dissimilarity ranking.

A single factor analysis of variance (ANOVA) was also employed for the analysis of the concentrations determined from the curve resolution analysis. Both a five class (HRC A, HRC B, HRC C, CV and WV) and a three class (in which the three HRC control samples are grouped together) ANOVA were investigated. The F values, a ratio of the mean squared variance between classes to the mean squared variance within classes, were used for comparison and are given in Table 1. These numbers can be rather deceiving, in that, the between class variance remains relatively unchanged when comparing a three class and five class ANOVA. However, the within class variance can be very different. This can be due to several reasons such as

Table 8.1: Peak statistics, SI and FR values

	Peak 1	**Peak 2**	**Peak 3**	**Peak 4**	**Peak 5**	**Peak 6**	**Peak 7**	**Peak 8**	**Peak 9**
Average relative concentration	13.84	51.11	26.94	16.74	21.48	3.11	14.23	3.71	9.35
Standard deviation	5.73	20.20	4.78	5.87	5.03	0.52	0.84	0.79	1.07
% RSD	41.39	39.53	17.76	35.06	23.39	16.76	5.92	21.23	11.42
Range	15.66	61.65	13.31	15.17	15.16	1.49	2.81	2.13	2.86
IQR	5.78	7.89	6.28	10.68	3.38	0.88	1.10	1.36	0.47
F, ANOVA 5 group	5125.75	23157.16	831.88	589.23	1959.82	52.66	15.46	58.95	276.82
F, ANOVA 3 group	182.78	382.89	43.21	6.75	134.68	21.09	1.59	4.08	602.76
SI	0.2485	0.3665	0.5009	0.5349	0.6318	0.6796	0.7381	0.8371	0.8429
FR 5 group	1.59×10^6	2.52×10^5	1.65×10^6	4.95×10^4	4.02×10^5	2.89×10^4	2.59×10^4	2.49×10^5	3.95×10^4
FR 3 group	6.81×10^5	1.65×10^5	4.14×10^5	7.14×10^4	3.45×10^5	1.57×10^4	4.54×10^4	1.31×10^5	9.19×10^4

compound degradation. Peak 4 exhibits the latter behavior, in that the concentrations of replicates HRC A are approximately half that of the concentrations of replicates HRC B and C (for unknown reasons). The within class variance is much higher for this sample when analyzed as one group, using a three class ANO VA (18.9) as opposed to three groups, using a five class ANOVA (0.20). In the case of peak 9, which is ranked as the most different by the three class ANOVA and the 6th most different using a five class ANOVA, the between class (15.0 and 15.81, three class and five class respectively) and within class (0.15 and 0.14) sum of squares values are very similar. With such a low between class sum of squares, the division by the degrees of freedom plays a large role in the final F value. Further discussion of ANOVA results for peak 9 are in section 8.3.4.

8.3.2 Threshold Determinations

To determine which peaks were to be excluded from further analysis, a means of determining a threshold value for similarity vs. dissimilarity was required for each of the compared methods (SI and FR). This determination requires some type of replicates to determine the within class variability using both methods. To accomplish this, the data was arranged to include only the three replicate HRC A samples, and analysis by either SI or FR was employed. This procedure was followed for the replicates of HRC B, HRC C, WV and CV. Because the SI method is not class-based, we were able to apply the algorithm to each of the five data sets and determine a minimum similarity index (minSI) value for each of the five different classes, see Table 8.2, for the respective minSI and corresponding peak numbers. The Fisher ratio method, as described earlier, uses the ratio of the class-to-class variance to the within class variance to

estimate similarities. Hence, the code was modified to apply only the within group portion to each class of data using $K = 3$ for the total number of samples and $Q = 1$ for the number of

Table 8.2: Minimum similarity index found for the replicate analysis of the 5 different samples studied.

Sample	**Minimum SI value/ Corresponding Peak #**	**Maximum FR value/ Corresponding Peak #**
HRC Control A	0.6848/ Peak 1	2.69×10^4/ Peak 1
HRC Control B	0.5062/ Peak 1	7.57×10^4/ Peak4
HRC Control C	0.6464/ Peak 7	4.37×10^4/ Peak 4
Ceplecha (CV)	0.7049/ Peak 1	1.58×10^4/ Peak 3
Winter (WV)	0.6303/ Peak 5	1.69×10^4/ Peak 5
Calculated Geographical Variability Threshold	0.6346	3.58×10^4

classes. While we realize this is not a valid statistical approach for the implementation of the Fisher ratio, we only claim to use this as a method to achieve a corresponding cut off value for inclusion/exclusion of peaks for further analysis. The result from this approach was verified using the Tukey (HSD) test. From this, we can conclude that any index values found when analyzing the geographical variability data that are 1) below the lowest minSI for the replicates (SI(HRC B) = 0.5062) in the case of the SI method or 2) above the highest maxFR for the replicates (FR(HRC B) = 7.57×10^4) in the case of the FR method, will be solely due to between class variability and not due to replicate variability. However, choosing the minSI or the maxFR as the threshold increases the probability of excluding peaks from further analysis that are truly significantly different between sample types, a type 2 error. If, on the other hand, we chose to use the greatest minSI (SI(CV) = 0.7049) or the lowest maximum FR (FR(CV) = 1.58×10^4) as a

threshold, we run the risk of targeting peaks for further analysis that are not significantly different, a type 1 error. As a tradeoff, we chose to use the average of the corresponding five index values from the replicate variability analyses as a threshold for the geographical variability (SIT = 0.6346, FRT = 3.58×10^4). As a side note, it should be mentioned that the above procedure for determining a SI threshold value, while suggested to the authors by Windig in a personal communication, was not performed in his experiments as he did not have replicate sample injections for each class.

To confirm the validity of the determined threshold values, two different statistical tests were investigated, an equivalence test (the reverse of a significance test) [124] and Tukey's HSD test (a significantly different test) [123], Tables 8.3 and 8.4, respectively. The equivalence test was used to determine if the means of the five different sample types are practically equivalent. JMP 8 software uses a two one-sided test (TOST) approach which requires an input of the difference considered to be practically zero since there is always some instrument variability. In order to validate the threshold values we used the IKSFA-ALS-ssel results as the true values for the concentrations of the nine evaluated peaks, and used the above tests applied to these values.

Because the nine HRC control samples should be most representative of within experiment variability, the total range for each of the determined concentrations of the nine peaks was calculated and the median of those values was used as the equivalence input criterion. The results of the equivalence test are found in Table 8.3. Peaks 6-9, are found to be equivalent between all sample classes while peaks 1-5 exhibit differing degrees of non equivalence with the equivalent classes, mainly between the HRC samples. These samples are expected to have some similarities since they are from the same batch of wine. All class comparisons for peak 2 were found to be not equivalent, peak 4 has nine non equivalent comparisons, peak 3 has eight non

equivalent comparisons, peak 5 has seven non equivalent comparisons and peak 1 has four non equivalent comparisons. The dissimilarity ranking for the equivalence test (Table 8.3) was

Table 8.3: Results of Equivalence Test where E shows equivalence between the 2 samples and NE showed no equivalence.

Sample Pairing	**SI[a]< 0.6346**					**SI[a]> 0.6346**			
	peak 2	**peak 4**	**peak 3**	**peak 5**	**peak 1**	**peak 6**	**peak 7**	**peak 8**	**peak 9**
HRC B/HRC A	NE	NE	NE	E	E	E	E	E	E
HRC C/HRC A	NE	NE	NE	E	E	E	E	E	E
HRC C/HRC B	NE	E	E	E	E	E	E	E	E
CV / HRC A	NE	NE	E	NE	E	E	E	E	E
CV / HRC B	NE	NE	NE	NE	E	E	E	E	E
CV / HRC C	NE	NE	NE	NE	E	E	E	E	E
WV/ HRC A	NE	NE	NE	NE	NE	E	E	E	E
WV/ HRC B	NE	NE	NE	NE	NE	E	E	E	E
WV/ HRC C	NE	NE	NE	NE	NE	E	E	E	E
WV/CV	NE	NE	NE	NE	NE	E	E	E	E

[a] the SIT (similarity index threshold value) = 0.6346

Table 8.4: Results of Tukey's (HSD) test where D shows a statistical difference between the 2 samples and ND shows no difference.

Sample Pairing	**FR[b]> 3.58 x 10^4**							**FR[b]< 3.58 x 10^4**	
	peak 1	**peak 2**	**peak 5**	**peak 3**	**peak 4**	**peak 8**	**peak 9**	**peak 6**	**peak 7**
HRC B/HRC A	D	D	D	D	D	D	ND	ND	ND
HRC C/HRC A	D	D	D	D	D	D	ND	ND	ND
HRC C/HRC B	D	D	D	ND	ND	ND	ND	ND	ND
CV / HRC A	D	D	D	D	D	D	ND	ND	ND
CV / HRC B	D	D	D	D	D	ND	ND	D	ND
CV / HRC C	D	D	D	D	D	ND	ND	D	ND
WV/ HRC A	D	D	D	D	D	ND	D	ND	ND
WV/ HRC B	D	D	D	D	D	D	D	D	ND
WV/ HRC C	D	D	D	D	D	D	D	ND	ND
WV/CV	D	D	D	D	D	D	D	D	ND

[a] the FRT (Fisher ratio index threshold value) = 3.58 x 10^4

determined to be peak 2 4 3 5 1 6 9 which corresponds very well to the SI method's ranking and threshold cut off such that peaks 6-9 are above the SIT = 0.6346.

However, the threshold value for the FR method only excludes peaks 6 and 7 with peak 9 being slightly above the cut off. The Tukey (HSD) test is a multiple comparisons (pairwise) test of the means for each class (the five different wine samples) and is an indicator of significant difference based on the standardized range statistic. It is calculated using the absolute value of the difference of two class means divided by the square root of the quantity of the within class mean square variance divided by the number of samples. The results for HSD test are also shown in Table 8.4. Peaks 1, 2 and 5 for all class comparisons were determined to be different. Peaks 3 and 4 are different for all comparisons except HCR C and HCR B. Six out the ten comparisons are significantly different for Peak 8. Peaks 6 and 9 are different for only four comparisons. Peak 7 is the only peak that is not significantly different for any of the class comparisons. Comparing the results in Table 8.4 for the peak rankings for the Tukey analysis, the five class ANOVA, and the five group FR analysis, Peaks 1, 5, 6 and 7 are all ranked the same, while peaks 4, 8 and 9 are clustered for each method around a rank of 5^{th}, 6^{th} or 7^{th} most dissimilar. Peaks 2 and 3 exchange rank order for the FR method as compared to the Tukey and five class ANOVA results predict such that Peak 3 becomes the most dissimilar in the FR analysis, Table 8.5. This deviation may be explained by the multiple, very small, overlapping peaks associated with peak 3 as compared to peak 2, making peak 3 more dissimilar based on chromatographic conditions and not simply on concentrations. The equivalence test clearly indicates that peaks 6-9 are equivalent, while the Tukey test results show that only peak 7 is not different with peaks 6 and 9 each having six means comparisons out of the ten that are classified as not statistically different. The use of the within class mean square variance, which is directly analogous to what

is done using an ANOVA analysis of the relative concentrations, and is also used in the FR calculations, helps to explain the better agreement of the Tukey test with the FR method and five class ANOVA than with that of the SI method.

Table 8.5: Peak rankings for the wine data for the SI and FR methods

Peak Ranking Order*									
SI	1	2	3	4	5	6	7	8	9
% RSD	1	2	4	5	8	3	6	9	7
Range	2	1	4	5	3	7	9	8	6
IQR	4	2	3	1	5	8	7	6	9
Equivalence test *	2	4	3	5	1	6	7	8	9
Tukey (HSD) Test *	1	2	5	3	4	8	6	9	7
ANOVA 5 group	2	1	5	3	4	9	8	6	7
ANOVA 3 group	9	2	1	5	3	6	4	8	7
FR 5 group	3	1	5	2	8	4	9	6	7
FR 3 group	1	3	5	2	8	9	4	7	6

*Peak rank is assigned in numerical order according to number of equivalent or different comparisons made for the appropriate statistical analysis; peaks listed in order of least similar to most similar from left to right. The gray boxes indicate peaks that are ranked identically so the order is arbitrary. The boxed peaks are ranked identically for both the 5 and 3 group FR analyses.

8.3.3 Geographical Variability (Similarity Index Method)

As mentioned earlier, in an effort to eliminate or at least minimize the effects of retention time shifting on the SI values, the data were globally aligned prior to the SI calculations. However, upon closer inspection of the globally aligned data, it was apparent that not all of the second dimension retention time shifting had been eliminated, and that several peaks also

exhibited a small degree of first dimension shifting. To aid in the determination of the effect retention time shifting has on the SI method, a comparison of with and without global alignment was performed. Before alignment the largest second dimension retention time shift is exhibited by peak 1 (13 data points equivalent to 0.325 seconds), peaks 2, 3 and 8 have a maximum second dimension retention time shift of 11 data points (0.275 seconds), peaks 4 and 5 have a maximum second retention time shift of 10 data points (0.25 seconds) and peaks 6, 7 and 9 have a shift of 9 data points (0.225 seconds). This order is significantly changed after alignment and the retention time shifts range from 8 data points (0.2 seconds) for peak 2 to 3 data points (0.075 seconds) for peak 3. Note that peak 3 prior to alignment shifts by 11 data points (the second greatest shift) but only by 3 data points after alignment (the smallest shift). This is significant to the SI value order change associated with retention time shifting, in that, without alignment the SI value order from most different (lowest SI value) to least different is peak 3 1 4 2 5 6 7 8 9. There are several things of importance to note here. First, the SI value does not directly correlate to the degree of retention time shifting either before or after alignment. Second, there is a retention time shift effect on the SI value peak order for the four most different peaks; however, the order does not change for peaks 5-9. This is significant, in that, the ranking of the peaks associated with a SI value greater than 0.6346 are not changed due to retention time shifting; and hence, those peaks are excluded from further analysis in either case, with or without alignment.

While the dissimilarity order of the equivalence test does not directly correspond to that of the dissimilarity order for the SI algorithm, these results correspond very nicely to the determined SIT of 0.6346. Peaks 6-9 (those found to be equivalent) have a SI greater than that of the SIT = 0.6346 and can therefore be excluded from further analysis as being too similar to require further consideration. From a quick inspection of the contour plot of only those peaks

with similarity index values less than that of the SIT, only twelve peaks are shown to exhibit concentration differences great enough to warrant further chemometric analysis; thus eliminating over half of the visually present peaks in the analyzed section of the chromatograms (originally ~30 peaks were noted, as discussed previously).

8.3.4 Geographical Variability (Fisher Ratio Method)

Because the FR method is a class-based calculation, we had the choice of either a five class or a three class analysis depending on if the three HRC samples are grouped as one class (since they should be very similar coming from the same batch) or as three different classes. The three class analysis sets $Q = 3$ so that the number of replicates is $n_q = 9$ for the HCR class and $n_q = 3$ for the CV and WV classes. As seen in Table 8.5, the dissimilarity order is relatively unaffected such that peaks 2, 5 and 8 (shaded area) remain in the same order while peaks 1 and 3, 4 and 9, 6 and 7 are simply swapped between the two analysis variations. Since there is not a significant difference in the dissimilarity order, a three class ANOVA analysis was performed using only the nine HRC samples, $ANOVA_{HRC}$. This was done to determine if these three within batch samples (HCR A, HCR B and HCR C) could be classed as not statistically different. The F critical value is 5.14 for the $ANOVA_{HRC}$ analysis and the calculated F values ranged from 0.3596 (peak 9) to 724.93 (peak 2) with the only peak having an FR value less than the F_{crit} value being peak 9. This leads to several conclusions. Other than peak 9, a statistical difference exists between the concentrations in the three HRC samples; and thus, these samples should be treated as three distinct classes. Also, as reported in Section 8.3.1, the three class ANOVA for the analysis of all five sample types, assigns peak 9 as the most dissimilar peak. However, for all of the other analysis methods, it is typically among the most similar peaks. The reason for this is now apparent. Since it is the only peak within this analysis where the three HRC samples are not

significantly different, the within class variance is very low leading to an artificially high *F* value in comparison to the other peaks that should statistically not be classed together but have been.

A comparison of the performance of the FR method with and without global alignment was also performed. The dissimilarity rank for the FR with alignment is reported in Table 8.5, while the dissimilarity rank prior to alignment is as follows: 2 1 5 8 4 6 9 3 7 from most dissimilar to most similar. Only peak 1 remains in the same dissimilarity position; all other peaks have been ranked differently. Peak 3 presents the greatest departure in dissimilarity order between the two analyses. Prior to alignment it is in the second highest retention time shifting group (11 data points) and is the next to the last most similar; while after alignment, it has the lowest retention time shift (3 data points) but is now the most dissimilar. This is rather counter intuitive and we have no explanation for the large change in dissimilarity.

The calculated threshold, FRT = 3.58×10^4, (Table 8.5) only excludes peak 6 in the FR three group analysis and excludes peaks 6 and 7 in the five group FR analysis. From a contour plot of only the data points that are above the FRT value, there are 17 peaks that will require further analysis. However, the contour plot does not show peaks 9 and 4, which are above the FRT and so should not be excluded from further analysis. A closer look at the raw data from the FR method reveals a total of 20 peaks that would need further investigation. This allows for the exclusion of approximately one-third of the peaks.

8.3.5 Simulated data analysis

The simulated data was used to answer several questions: (1) Does the background signal affect the SI or FR values and/or ranks? (2) Is there an effect on the algorithms compared due to chromatographic conditions such as retention time alignment, peak overlap and spectral differences? (3) Using simulated data, can we determine what the true SI or FR values/ranks of

the nine peaks are? This would allow for better evaluation of the performance of the compared algorithms. To address the first question in regards to the effect the mobile phase background may have on the algorithms, four varying combinations of background components were compared and both SI and FR methods were applied to each background varied simulated data set. A control mixture (consisting of known indoles) was injected at the beginning (background 1) and the end (background 2) of the wine sample run. The corresponding section (3.85 to 10.45 seconds second retention time dimension by 3.5 to 14.35 minutes first retention time dimension) of the two standard mixture injections to model the background variations in the simulated data. The four simulated data sets utilized (a) background 1 for all ten sample injections, (b) 100 % background 1 for injection 1 and evenly decreased this percentage to 0 for injection 10 while increasing background 2 from 0 % for injection 1 to 100 % for injection 10, (c) 100 % background 1 for injection 1 to 0 % background 1 for injection 10 and 0 % background 2 for injection 1 to 50 % for injection 10, and (d) same ratio as for (b) but the overall background intensity was doubled.

For background combinations a, b and c using the SI method, the peak order was 1 2 4 5 3 8 6 7 9 with the only change in SI order being peaks 1 and 2 which simply exchanged positions in background combination b, Table 8.6. This leads to two significant points. The first, that while the dissimilarity peak rank of the simulated data is not the same as that of the experimental data, the same peaks will be excluded from further analysis in both data sets. Second, that a changing background over the course of a LC × LC run has very little effect on the SI method. Background combination d gave a different order: 2 1 8 6 4 9 3 5 7 from the above 3 combinations and from the experimental data. This is most likely due to a decrease in the signal to backgroud ratio and has the deleterious effect of including/excluding inappropriate peaks for

further analysis. The FR method placed peak 2 as the most different for all background combinations, peak 1 second most different for 3 out of the 4 combinations, peak 5 third most different for 3 out of the 4 combinations, but after that the order changed for each background combination. Varying the background leads to less reproducible results for the FR method resulted as compared to the SI method.

Table 8.6: Peak rankings for the simulated data consisting of four different background contributions for the SI and FR methods.

	Similarity Index method comparison of backgrounds								
Background a[1]	1	2	4	5	3	8	6	7	9
Background b[2]	2	1	4	5	3	8	6	7	9
Background c[3]	1	2	4	5	3	8	6	7	9
Background d[4]	2	1	8	6	4	9	3	5	7
	Fisher ratio method comparison of backgrounds								
Background a	2	1	3	6	4	9	7	5	8
Background b	2	9	5	1	3	6	7	4	8
Background c	2	1	5	9	3	7	8	4	9
Background d	2	1	5	3	9	4	6	7	8

[1] background 1 for all ten sample injections.
[2] 100 % background 1 for injection 1 and evenly decreased this percentage to 0 for injection 10 while increasing background 2 from 0 % for injection 1 to 100 % for injection 10.
[3] 100 % background 1 for injection 1 to 0 % background 1 for injection 10 and 0 % background 2 for injection 1 to 50 % for injection 10.
[4] same ratio as for background b but the overall background intensity was doubled.

A second simulated data set was made to determine the effect of spectral differences on the index values of the SI and FR methods. To that end, the simulated data was created such that there is no background component, the concentrations of two peaks are the same, and there is no retention time shifting or overlapping between the two peaks. Each peak was assigned very different normalized spectra as the only contributing difference to the simulated data. The SI was determined to be 0.3447 for both of the peaks indicating that the two peaks are essentially the same with regard to all contributing chromatographic factors; hence, spectral differences do not contribute to the SI value. Because of this, any difference in the SI values for the

experimental data and for the simulated nine peak data without a background will be due to variations in the chromatographic conditions or due to real concentration variations. Using the simulated nine peak data without background, the true ranking order was determined to be 1 2 4 5 3 8 6 7 9 and the SI value range for the nine peak data was 0.2817 to 0.9074. The FR method, however, yielded quite a different result. Each of the two simulated peaks with different spectra have a different FR value ($2.14 \times 10^6$and 2.39×10^6). This indicates that the different spectra associated with each compound will have an effect on the FR values. Mohler *et al*. note that because the Fisher ratio method is based on signal intensities, preprocessing of the raw data to normalize and to baseline correct would be appropriate. In the case of this work, a blank was unavailable for background subtraction of the experimental data, and the FR algorithm did not allow for implementation without a background of the nine peak simulated data. Therefore the FR value is not totally dependent on concentration, and thus a true rank order cannot be determined using this method. It may also explain some of the divergent results between the ANOVA analysis and the FR method because the ANOVA results are truly based on only peak concentrations.

8.4 Conclusions

These results show that both the SI and FR methods can be used as a rapid screening method for LC × LC-DAD analyses of complex biological data. Both methods were able to locate areas of the chromatograms on which to focus further quantitative analyses of the wine samples. The advantage of implementing either of these methods prior to quantitative analysis is the dramatic reduction in the data analysis time. This reduction in analysis time occurs because only the significantly different peaks are resolved and quantified as opposed to the resolution and quantification of all peaks present in the entire chromatogram and then using those quantification

results to determine to the peaks of significant differences between samples by means of concentration differences. The SI method has several advantages over the FR method. The similarity index method is based on a correlation coefficient analysis and is not affected by the different spectra associated with the differing compounds. Also, no prior class knowledge of the samples is required. This allows for the rapid and efficient screening of samples of unknown origin that may belong to multiple different classes. Retention time shifting in both the first and second dimensions, along with background variability and overlapped peaks were all chromatographic conditions that were shown to affect the SI values. These chromatographic conditions also affect the FR method, which is ANOVA based. This method is also affected by the different spectra associated with individual compounds and requires class knowledge of the analyzed samples. The determined true SI ranking from the simulated data is shown to be very comparable with the peak rankings assigned to the nine peaks in the wine geographical variability such that peaks 1, 2 and 9 are ranked correctly and peaks 3 and 4, 5 and 6 are shifted by only one rank position. Due to the effect the spectra have on the FR values, the true rank of the FR analysis could not be determined. The equivalence test verifies the use of the determined threshold for the SI method, while Tukey's (HSD) test confirmed the choice of threshold used with the FR method. There are several advantages to the determination and use of the threshold values. The calculation is quick and easy; and because of this, the user can vary this value to suit the individual analysis needs.

Chapter 9: Conclusions and Future Work

Advances in comprehensive two-dimensional liquid chromatography have brought with them several significant accomplishments. One of the most important is the ability to chromatographically separate very complex samples in a relatively short time span as compared to less than a decade ago when run times per sample could take hours or days. The possible information that can be gleaned from such samples is staggering. This is very exciting especially for researchers in the "-omics" fields. However, the ability to separate very complex samples has led directly to very complex data sets that are not easily analyzed by current chemometric techniques. The need for appropriate software that is not only user friendly but that also provides accurate and precise results is of the utmost importance. This work sought to advance chemometric applications with respect to resolution of rank deficient four-way data. This led to an investigation of quantification methods for four-way data with the additional aim of further understanding the chromatographic factors that affect peak quantification of data arising from LC × LC-DAD data. While the developed IKSFA-ALS method followed by manual baseline integration is shown to provide both accurate and precise quantification results, it is rather tedious and cumbersome.

Because the goal of LC × LC analyses is sometimes the resolution and quantification of only a few target compounds out of very complex samples, total chemometric analysis time

would be better spent on only the compounds that are of interest. This is a perfectly reasonable statement when the target compound/compounds is/are known *a priori*. This becomes an overwhelming issue in very complex samples, however, when the target compound is not known in advance of either the chromatographic separation or the chemometric data analysis. To alleviate this issue, rapid screening methods, that had previously been applied to other types of data, were investigate for their applicability to four-way LC × LC data with known retention time shifting issues.

9.1 Goal of Resolution

The goal of developing a chemometric method that was unaffected by retention time shifting and capable of handling rank deficient, four-way LC × LC-DAD data, was achieved by the IKSFA-ALS method. Using this chemometric technique, over fifty peaks were resolved (only eighteen of those were observable from a contour plot before chemometric resolution) from a section of urine control data consisting of fourteen replicate samples. For example, one observable peak was chemometrically resolved to reveal two other underlying compounds with very similar first and second retention times but associated with very different spectra. The method was also shown to place the background signal into separate components, effectively removing it from components containing analyte signals to be quantified.

A comparison of the chemometric resolving and quantitative power of PARAFAC versus the developed MCR-ALS method was performed. This analysis was performed to determine the extent of the deleterious effect the lack of multilinearity, due to retention time shifting in both the first and second retention time dimensions, may have on peak resolution of the two methods. In other words, can methods that require multilinearity be employed without alignment or other

preprocessing of the data without a loss in accuracy and precision of quantification? (the quantitative results are discussed below). The chemometric method described in this research project found and resolved six additional peaks not detected by the PARAFAC-ALS method. Additionally, the IKSFA-ALS method resolved several peaks that the PARAFAC-ALS method was able to detect but not resolve due to ^{1}D retention time shifting.

9.2 Goal of Quantification

Once a suitable chemometric resolution method was developed, it was important to investigate possible quantification methods for the resolved peaks. When this research project began, Stoll *et al.* had shown quantification of compounds arising from LC × LC separations to be less precise as compared to quantification of compounds arising from 1D-LC separations [83]. An investigation into the possible reasons for the lack of precision of LC × LC separated compounds as compared to 1D-LC was deemed essential to this work and was undertaken. Simultaneous investigation into multiple different peak area/volume determination approaches was completed to determine advantages and disadvantages of these quantification methods. The standards mixture data was utilized for this purpose and the precision of quantification was employed as a measure of success of the investigated approaches when applied to both raw data and to chemometrically resolved data. The total summation method was only applicable to the quantification of peaks after chemometric background removal and resolution. The manual baseline method was found to yield overall better precision of quantification of both the raw data (twice as precise as the commercially available LCImage software) and chemometrically resolved data (six times more precise than LCImage software and four times better than the total summation method). There was no improvement in precision of the LCImage software of the resolved data versus the raw data. This is possibly due to the manner in which the software

determines a peak baseline. However, using the manual baseline method, there was a 2.5 fold increase in precision of the IKSFA-ALS analyzed data as compared to the precision of the raw data. Precision of quantification was also compared for the maize data analysis using the IKSFA-ALS and PARAFAC-ALS methods. The IKSFA-ALS method resulted in 3.8 to 10.5 fold improvement in precision for five out of the six analyzed peaks from the maize data.

Accuracy of quantification of the IKSFA-ALS method was investigated using the data derived from WWTPE samples where the goal was to resolve the phenytoin peak from the chromatographically overlapped interferent. This data included both a calibration set and a standard addition set of experimental injections. The severe overlap of the interferent with the analyte rendered quantification of the phenytoin peak impossible for the sLC × LC raw data without chemometric resolution. An additional constraint, sample selectivity, was employed with the IKSFA-ALS method. The concentration of the phenytoin in the unspiked WWTPE was determined to be 42 ± 1 ng/L using the standard addition method which corresponds very well with the result obtained from the 2D-LC-MS/MS reference method. The external calibration method results were slightly lower, 36 ± 1 ng/L. The cause of this inconsistency was shown, by statistical analyses, to be the result of matrix effects in the WWTPE samples. The % RSD of precision of the duplicate injections of the DI water and WWTPE samples with the implementation of the sample selectivity constraint results in a six-fold improvement as compared to the chemometric results when only the spectra selectivity constraint was utilized.

9.3 Goal of Rapid Screening

A major drawback to the IKSFA-ALS method followed by manual baseline integration is the lack of automation requiring tedious user intervention. If the goal of a given data analysis is

to chemometrically locate, resolve and quantify as many compounds in a very complex sample where large dynamic ranges in concentrations exist, this method has been shown to achieve that goal, if in a somewhat time consuming manner. However, often the goal is to locate, resolve and then quantify only compounds in the said complex sample that show significant changes in concentrations between varying sample types. In the case of human metabolomic samples (such as urine, blood etc…), hundreds of the compounds present may have no significant concentration differences and therefore, do not require resolution and quantification. Only a handful of the metabolites in the sample may be indicative of a disease state, toxicity or drug efficacy. In the case of wine analysis, these concentration differences may be indicative of geographical, varietal or fermentation variability between samples. With the drawbacks of the developed method and the goal of resolution and quantification of only unknown target compounds, a rapid screening method to locate the significant compounds before resolution and quantification was a reasonable next step. Two methods were compared, the Similarity Index method and the Fisher ratio method, for their applicability to four-way data. It was important that the only contributing factor to the screening method be that of the differences in concentration of the peaks from sample to sample. Hence the method needed to be robust against retention time shifting and spectral differences. Both the SI and FR methods were found to help in the identification of peaks having significant concentration differences, which would then be subjected to further analysis. The SI method was less affected by retention time shifting than that of the FR method and was also completely independent of spectral information.

9.4 Future Work

This work led to the development and implementation of a new unimodality constraint for the MCR-ALS algorithm that allows for the application of the unimodality condition in both

the first and second retention time dimensions [105]. Unlike traditional unimodality constraints where the lesser signal is essentially eliminated from the resolved component, this approach in the implementation of the unimodality constraint increases the number of components and then assigns the newly created unimodal peak to the new component. While this is an improvement, there is yet another step to be taken. The current method requires dividing the co-eluting peaks with a vertical or horizontal line. Unfortunately, this has an adverse effect on quantification, especially if the two peaks have very large concentration differences. What may be a better solution, would be to fit the two newly unimodally separated peaks to their own Gaussian models so that the "cut off" portions of each peak may be more appropriately quantified.

As has been discussed, the lack of multilinearity is a limiting factor when choosing a chemometric method. The lack of alignment algorithms for LC × LC data when this work began, led us to the use of MCR-ALS that only required the data to by bilinear. Since this time, further research within the Rutan group focused on a comparison of the following five interpolation algorithms: linear interpolation followed by cross correlation, piecewise cubic Hermite interpolating polynomial, cubic spline, Fourier zero-filling, and Gaussian fitting [125]. From this work, Allen *et al.* developed a semi-automated alignment method where the final step was a four-way PARAFAC analysis for resolution and quantification [126]. There is, however, some debate among the MCR-ALS/ PARAFAC communities about which is the better approach. Currently both methods have their unique advantages and disadvantages. With that said, I do not believe there is an absolute answer to the current debate. At present, if the goal of the analysis is precision and accuracy of quantification or resolution of low concentration compounds and time is not of importance, than MCR-ALS is an appropriate technique. If on the other hand, time is of the essence and accuracy and precision of quantification are of secondary importance,

PARAFAC with pre-alignment offers a clear advantage in the reduced analysis time and minimal user intervention required to arrive at a reasonable result.

Further improvements in the area of quantification both from the aspect of the chromatography conditions and from the development of more automated algorithms will broaden the applicability of the LC × LC analytical technique. The ultimate goal is to achieve complete peak resolution and high precision and accuracy of quantification in a timely manner with little to no subjective user intervention. This will require that the algorithm is able to correctly determine the number of components to be used, implement appropriate constraints and arrive at concentrations for the compounds in the sample. Future chromatographic and chemometric collaborations have the best chance of successfully addressing these issues that must be resolved before a truly user friendly software package can be designed.

References

[1] H.P. Bailey, S.C. Rutan, Chemometric Resolution and Quantification of Four-Way Data Arising from Comprehensive 2D-LC-DAD Analysis of Human Urine, Chemom. Intell. Lab. Sys., 106 (2011) 131-141.

[2] S.E.G. Porter, D.R. Stoll, S.C. Rutan, P.W. Carr, J.D. Cohen, Analysis of Four-Way 2D LC-DAD: Application to Metabolomics, Anal. Chem., 78 (2006) 5559-5569.

[3] D.R. Stoll, X. Li, X. Wang, P.W. Carr, S.E.G. Porter, S.C. Rutan, Fast, Comprehensive Two-Dimensional Liquid Chromatography, J. Chromatogr. A, 1168 (2007) 3-43.

[4] S. Wold, Chemometrics; What Do We Mean with it, and What Do We Want from it?, Chemom. Intell. lab. Sys., 30 (1995) 109-115.

[5] N.-P.V. Nielson, J.M. Carstensen, J. Smedsgaard, Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping, J. Chromatogr. A, 805 (1998) 17-35.

[6] B. Walczak, W. Wu, Fuzzy Warping of Chromatograms, Chemom. Intell. Lab. Sys., 77 (2005) 173-180.

[7] D. Bylund, R. Danielsson, G. Malmquist, K.E. Markides, Chromatographic Alignment by Warping and Dynamic Programming as a Pre-Processing Tool for PARAFAC Modeling of Liquid Chromatography-Mass Spectrometry Data, J. Chromatogr. A, 961 (2002) 237-244.

[8] M. Otto, Chemometrics, Wiley-VCH Verlag Gmbh & Co., Weinheim, 2007.

[9] T. Skov, F. van der Berg, G. Tomasi, R. Bro, Automated Alignment of Chromatographic Data, J. Chemom., 20 (2006) 484-497.

[10] R. Bro, PARAFAC. Tutorial and Applications, Chemom. Intell. Lab. Sys., 38 (1997) 149-171.

[11] A. De Juan, S.C. Rutan, R. Tauler, D.L. Massart, Comparison Between the Direct Trilinear Decomposition and the Multivariate Resolution-Alternating Least Squares Methods for the Resolution of Three-Way Data Sets, Chemom. Intell. Lab. Sys., 40 (1998) 19-32.

[12] H.P. Bailey, S.C. Rutan, D.R. Stoll, Chemometric Analysis of Targeted 3DLC-DAD Data for Accurate and Precise Quantification of Phenytoin in Wastewater Samples, J. Sep. Sci., 35 (2012) 1837-1843.

[13] H.P. Bailey, S.C. Rutan, Factors that Affect Quantification of Diode Array Data in Comprehensive Two-Dimensional Liquid Chromatography Using Chemometric Data Analysis, J. Chromatogr. A, 1218 (2011) 8411-8422.

[14] H.P. Bailey, S.C. Rutan, Comparison of Chemometric Methods for the Screening of Comprehensive Two-Dimensional Liquid Chromatographic Analysis of Wine, Anal. Chim. Acta, submitted July (2012).

[15] Y. Zhang, Chromatographic Selectivity and Hyper-Cosslined Liquid Chromatography Stationary Phases. Ph.D. Dissertation Department of Chemistry, University of Minnesota, Minnesota, 2010.

[16] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, Recent Advances in Comprehensive Two-Dimensional Separations with Chemometrics, J. Chromatogr. A, 1184 (2008) 341-352.

[17] M.T. Cantwell, S.E.G. Porter, S.C. Rutan, Evaluation of the Multivariate Selectivity of Multi-Way LC Methods, J. Chemom., 21 (2007) 335-345.

[18] P. Jandera, T. Hajek, P. Cesla, Comparison of Various Second-Dimension Gradient types in Comprehensive Two-Dimensional Liquid Chromatography, J. Sep. Sci., 33 (2010) 1382-1397.

[19] H. Malerod, E. Lundanes, T. Greibrokk, Recent Advances in On-Line Multidimensional Liquid Chromatography, Anal. Methods, 2 (2010) 110-122.

[20] S. Eeltink, S. Dolman, G. Vivó-Truyols, P. Scheonmakers, R. Swart, M. Ursem, G. Desmet, Selection of Column Dimensions and Gradient Conditions to Maximize the Peak-Production Rate in Comprehensive Off-Line Two-Dimensional Liquid Chromatography Using Monolithic Columns, Anal. Chem., 82 (2010) 7015-7020.

[21] T. Ikegami, T. Hara, H. Kimura, H. Kobayashi, K. Hosoya, K. Cabrera, N. Tanaka, Two-Dimensional Reversed-Phase Liquid Chromatography Using Two Monolithic Silica C18 Columns and Different Mobile Phase Modifiers in the Two Dimensions, J. Chromatogr. A, 1106 (2006) 112-117.

[22] S.E. Reichenbach, P.W. Carr, D.R. Stoll, Q. Tao, Smart Templates for Peak Pattern Matching with Comprehensive Two-Dimensional Liquid Chromatography, J. Chromatogr. A, 1216 (2009) 3458-3466.

[23] A.L. Huidobro, P. Pruim, P.J. Schoenmakers, C. Barbas, Ultra Rapid Liquid Chromatography as a Second Dimension in a Comprehensive Two-Dimensional Method for the Screening of Pharmaceutical Samples in Stability and Stress Studies, J. Chromatogr. A, 119 (2008) 182-190.

[24] A.J. Alexander, M. Lianjia, Comprehensive Two-Dimensional Liquid Chromatography Separations of Pharmaceutical Samples Using Dual Fused-Core Columns in the Second Dimension, J. Chromatogr. A, 1216 (2009) 1338-1345.

[25] D.R. Stoll, J.D. Cohen, P.W. Carr, Fast, Comprehensive Online Two-Dimensional High Performance Liquid Chromatography Through the use of High Temperature Ultra-Fast Gradient Elution Reversed Phase Liquid Chromatography, J. Chromatogr. A, 1122 (2006) 123-137.

[26] L.W. Potts, D.R. Stoll, X. Li, P.W. Carr, The Impact of Sampling Time on Peak Capacity and Analysis Speed in On-Line Comprehensive Two-Dimensional Liquid Chromatography, J. Chromatogr. A, 1217 (2010) 5700-5709.

[27] S.R. Groskreutz, M.M. Swenson, L.B. Secor, D.R. Stoll, Selective Comprehensive Mult-Dimensional Separation for Resolution Enhancement in High Performance Liquid Chromatography. Part I: Principles and Instrumentation, J. Chromatogr. A, 1228 (2012) 31-40.

[28] C.F. Poole, The Essence of Chromatography, Elsevier, Amsterdam, 2003.

[29] T. Teutenberg, High-Temperature Liquid Chromatography A User's Guide for Method Development, RSC Publishing, Cambridge, 2010.

[30] J.C. Giddings, Two-Dimensional Separations: Concepts and Premise, Anal. Chem., 56 (1984) 1258A.

[31] P.H. Farrell, High Resolution Two-Dimensional Electrophoresis of Proteins, J. Biol. Chem., 250 (1975) 4007.

[32] F. Erni, R.W. Frei, Two-Dimensional Column Liquid Chromatographic Technique for Resolution of Complex Mixtures, J. Chromatogr. A, 149 (1978) 561-569.

[33] G. Guiochon, N. Marchetti, K. Mriziq, R.A. Shalliker, Implementations of Two-Dimensional Liquid Chromatography, J. Chromatogr. A, 1189 (2008) 109-168.

[34] R.E. Murphy, M.R. Shure, J.P. Foley, Effect of Sampling Rate on Resolution in Comprehensive Two-Dimensional Liquid Chromatography, Anal. Chem., 70 (1998) 1585-1594.

[35] E. Grushka, N. Grinberg, Advances in Chromatography, in: P.W. Carr, J.M. Davis, S.C. Rutan (Eds.) Online Comprehensive Multidimensional Liquid Chromatography, CRS Press Taylor and Francis Group, Boca Raton, Florida, 2012.

[36] S.C. Rutan, J.M. Davis, P.W. Carr, Fractional Coverage Metrics Based on Ecological Home Range for Calculation of the Effective Peak Capacity in Comprehensive Two-Dimensional Separations, J. Chromatogr. A, 1255 (2012) 267-276.

[37] C.G. Fraga, C.A. Bruckner, R.E. Synovec, Increasing the Number of Analyzable Peaks in Comprehensive Two-Dimensional Separations Through Chemometrics, Anal. Chem., 73 (2001) 675-683.

[38] J.M. Davis, D.R. Stoll, P.W. Carr, Effect of First-Dimension Undersampling on Effective Peak Capacity in ComprehensiveTwo-Dimensional Separations, Anal. Chem., 80 (2008) 461-473.

[39] J.M. Davis, S.C. Rutan, P.W. Carr, Relationship Between Selectivity and Average Resolution in Comprehensive Two-Dimensional Separations with Spectroscopic Detection, J. Chromatogr. A, 1218 (2011) 5819-5828.

[40] D.J. Crockford, J.C. Lindon, O. Cloarec, R.S. Plumb, S.J. Bruce, S. Zirah, P. Rainville, C.L. Stumpf, K. Johnson, E. Holmes, J.K. Nicholson, Statistical Search Space Reduction and Two-Dimensional Data Display Approaches for UPLC-MS in Biomarker Discovery and Pathway Analysis, Anal. Chem., 78 (2006) 4398-4408.

[41] R. Kaddurah-Daouk, B.S. Kristal, R.M. Weinshilboum, Metobolomics: a Global Biochemical Approach to Drug Response and Disease, Annu. Rev. Pharmacol. Toxicol., 48 (2008) 653-683.

[42] C.G. Fraga, C.A. Corley, The Chemometric Resolution and Quantification of Overlapped Peaks from Comprehensive Two-Dimensional Liquid Chromatography, J. Chromatogr. A, 1096 (2005) 40-49.

[43] J. Pol, B. Hohnova, M. Jussila, T. Hyötyläinen, Comprehensive Two-Dimensional Liquid Chromatography-Time-of-Flight Mass Spectrometry in the Analysis of Acidic Compounds in Atmospheric Aerosols, J. Chromatogr. A, 1130 (2006) 64-71.

[44] M. Kivilompolo, T. Hyötyläinen, Comprehensive Two-Dimensional Liquid Chromatography in Analysis of Lamiaceae Herbs: Characterisation and Quantification of Antioxidant Phenolic Acids, J. Chromatogr. A, 1145 (2007) 155-164.

[45] M. Kivilompolo, V. Oburka, T. Hyötyläinen, Comprehensive Two-Dimensional Liquid Chromatograpy in the Analysis of Antioxidant Phenolic Compounds in Wines and Juices, Anal. Bioanal. Chem., 391 (2008) 373-380.

[46] L. Mondello, M. Herrero, T. Kumm, P. Dugo, H. Cortes, G. Dugo, Quantification in Comprehensive Two-Dimensional Liquid Chromatography, Anal. Chem., 80 (2008) 5418-5424.

[47] M. Kallio, T. Hyötyläinen, Quantitative Aspects in Comprehensive Two-Dimesnional Gas Chromatography, J. Chromatogr. A, 1148 (2007) 228-235.

[48] S.E. Reichenbach, Quantification in Comprehensive Two-Dimensional Liquid Chromatography, Anal. Chem., 81 (2009) 5099-5101.

[49] G. Vivó-Truyols, H.-G. Janssen, Probabiliy of Failure of the Watershed Algorithm for Peak Detection in Comprehensive Two-Dimensional Chromatography, J. Chromatogr. A, 1217 (2010) 1375-1385.

[50] D. Thekkudan, S.C. Rutan, P.W. Carr, A Study of the Precision and Accuracy of Peak Quantification in Comprehensive Liquid Chromatography in Time, J. Chromatogr. A, 1217 (2010) 4313-4327.

[51] E. Bezemer, S.C. Rutan, Multivariate Curve Resolution with Non-Linear Fitting of Kinetic Profiles, Chemom. Intell. Lab. Sys., 59 (2001) 19-31.

[52] L.A. Lopez, S.C. Rutan, Comparison of Methods for Characterization of Reversed-Phase Liquid Chromatographic Selectivity, J. Chromatogr. A, 965 (2002) 301-314.

[53] E.R. Malinowski, Factor Analysis in Chemistry, 2nd ed., John Wiley & Sons, Inc., New York, 1991.

[54] A. Bogomolov, M. McBrien, Mutual Peak Matching in a Series of HPLC-DAD Mixture Analyses, Anal. Chim. Acta, 490 (2003) 41-58.

[55] P.V. van Zomeren, H. Darwinkel, P.M.J. Coenegracht, G.J. de Jong, Comparison of Several Curve Resolution Methods for Drug Impurity Profiling Using HPLC-DAD, Anal. Chim. Acta, 487 (2003) 155-170.

[56] K.J. Schostack, E.R. Malinowski, Preferred Set Selection by Iterative Key Set Factor Analysis, Chemom. Intell. Lab. Sys., 6 (1989) 21-29.

[57] E.R. Malinowski, Automatic Window Factor Analysis- a More Efficient Method for Determining Concentration Profiles from Evolutionary Spectra, J. Chemom., 10 (1996) 273-279.

[58] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, J.P. Huvenne, Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations, Journal of Chemical Information and Computer Sciences, 43 (2003) 2057-2067.

[59] W. Windig, The Use of Second-Derivative Spectra for Pure-Variable Based Self-Modeling Mixture Analysis Techniques, Chemom. Intell. Lab. Sys., 23 (1994) 71-86.

[60] E. Bezemer, S.C. Rutan, Analysis of Three- and Four-Way Data Using Multivariate Curve Resolution-Alternating Least Squares with Global Multi-Way Kinetic Fitting, Chemom. Intell. Lab. Sys., 81 (2006) 82-93.

[61] N.M. Faber, R. Bro, P.K. Hopke, Recent Developments in CANDECOMP/PARAFAC Algorithms: a Critical Review, Chemom. Intell. lab. Sys., 65 (2003) 119-137.

[62] G. Tomasi, F. van den Berg, C. Andersson, Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data, J. Chemom., 18 (2004) 231-241.

[63] K.M. Pierce, L.F. Wood, B.W. Wright, R.E. Synovec, A Comprehensive Two-Dimensional Retention Time Alignment Algorithm to Enhance Chemometric Analysis of Comprehensive Two-Dimensional Separation Data, Anal. Chem., 77 (2005) 7735-7743.

[64] R.A. Harshman, PARAFAC- Explanatory Factor Analysis Procedure, J. Acoust. Soc. Am., 50 (1971) 117.

[65] J.D. Carroll, J.J. Chang, Analysis of Individual Differences in Multidimensional Scaling via N-Way Generalization of Eckart-Young Decomposition, Psychometrika, 35 (1970) 282.

[66] J.A. Arancibia, A.C. Olivieri, D.B. Gil, A.E. Mansilla, I. Duran-Meras, A.M. de la Pena, Trilinear Least-Squares and Unfolded-PLS Coupled to Residual Trilinearization: New Chemometric Tools for the Analysis of Four-Way Instrumental Data, Chemom. Intell. Lab. Sys., 80 (2006) 77-86.

[67] H. Idborg, P.-O. Edlund, S.P. Jacobson, Multivariate Approaches for Efficient Detection of Potential Metabolites from Liquid Chromatography/Mass Spectrometry Data, Rapid Commun. Mass Spectrom., 18 (2004) 944-954.

[68] H.L. Wu, M. Shibukawa, K. Oguma, An Alternating Trilinear Decomposition Algorithm with Application to Calibration of HPLC-DAD for Simultaneous Determination of Overlapped Chlorinated Aromatic Hydrocarbons, J. Chemom., 12 (1998) 1-26.

[69] J.H. Jiang, H.L. Wu, Y. Li, R.Q. Yu, Three-Way Data Resolution by Alternating Slice-Wise Diagonalization (ASD) Method, J. Chemom., 14 (2000) 15-36.

[70] J.M.F. ten Berge, A.K. Smilde, Non-Triviality and Identification of a Constrained Tucker3 Analysis, J. Chemom., 16 (2002) 609-612.

[71] R. Pardo, B.A. Helena, C. Cazurro, C. Guerra, L. Deban, C.M. Guerra, M. Vega, Application of Two- and Three-Way Principal Component Analysis to the Interpretation of Chemical Fractionation Results Obtained by the use of the B.C.R. Procedure, Anal. Chim. Acta, 523 (2004) 125-132.

[72] J.J. Jansen, R. Bro, H.C.J. Hoefsloot, F.W.J. van den Berg, J.A. Westerhuis, A.K. Smilde, PARAFASCA: ASCA Combined with PARAFAC for the Analysis of Metabolic Fingerprinting Data, J. Chemom., 22 (2008) 114-121.

[73] C.A. Andersson, R. Bro, The N-way Toolbox for MATLAB, Chemom. Intell. Lab. Sys., 52 (2000) 1-4.

[74] M. Daszykowski, W. Wu, A.W. Nicholls, R.J. Ball, T. Czekaj, B. Walczak, Identifying Potential Biomarkers in LC-MS Data, J. Chemom., 21 (2007) 292-302.

[75] E. Gebel, Mini-Metabolomics, Anal. Chem., 80 (2008) 3947.

[76] F. Gan, Q.S. Xu, Y.Z. Liang, Two Novel Procedures for Automatic Resolution of Two-Way Data from Coupled Chromatography, Analyst, 126 (2001) 161-168.

[77] J.C. Lindon, E. Holmes, J.K. Nicholson, So What's the Deal with Metabonomics?, Anal. Chem., (2003) 385A.

[78] J.C. Lindon , E. Holmes, M.E. Bollard, E.G. Stanley, J.K. Nicholson, Metabonomics Technologies and Their Applications in Physiological Monitoring, Drug Safety Assessment and Disease Diagnosis
Biomarkers, 9 (2004) 1-31.

[79] B. Shrestha, Y. Li, A. Vertes, Rapid Analysis of Pharmaceuticals and Excreted Xenobiotic and Endogenous Metabolites with Atmospheric Pressure Infrared MALDI Mass Spectrometry, Metabolomics, 4 (2008) 297-311.

[80] E.Y. Xu, W.H. Schaefer, Q.W. Xu, Metabolomics in Pharmaceutical Research and Dvelopment: Metabolites, Mechanisms and Pathways, Curr. Opin. Drug Discov. Dev., 12 (2009) 40-52.

[81] S. Smith, H. Burden, R. Persad, K. Whittington, B. de Lacy Costello, N.M. Ratcliffe, C.S. Probert, A Comparative Study of the Analysis of Human Urine Headspace using Gas Chromatography-Mass Spectrometry, J. Breath Res., 2 (2008) 1-10.

[82] K.K. Pasikanti, P.C. Ho, E.C.Y. Chan, Development and Validation of a Gas Chromatography/Mass Spectrometry Metabonomics Platform for the Global Profiling of Urinary Metabolites, Rapid Commun. Mass Spectrom., 22 (2008) 2984-2992.

[83] D.R. Stoll, X. Wang, P.W. Carr, Comparison of the Practical Resolving Power of One- and Two-Dimensional HPLC Analysis of Metabolomic Samples, Anal. Chem., 80 (2008) 268-278.

[84] Y. Zhang, L. Hao, P.W. Carr, Silica-based, Hyper-crosslinked Acid Stable Stationary Phases for High Performance Liquid Chromatography, J. Chromatogr. A, 1228 (2012) 110-124.

[85] I.J. Kosir, J. Kidric, Use of Nuclear Magnetic Resonance Spectroscopy in Wine Analysis: Determination of Minor Compounds, Anal. Chim. Acta, 458 (2002) 77-84.

[86] B. Mendes, J. Goncalves, J.S. Camara, Effectiveness of High-Throughput Miniaturized Sorbent- and Solid Phase Microextraction Techniques Combined with Gas Chromatography-Mass Spectrometry Analysis for a Rapid Screening of Volatile Semi-Volitle Composition of Wines-A Comparative Study, Talanta, 88 (2012) 79-94.

[87] A. Cuadros-Inostroza, P. Giavalisco, J. Hummel, A. Eckardt, L. Willmitzer, H. Pena-Cortes, Discrimination of Wine Attributes by Metabolome Analysis, Anal. Chem., 82 (2010) 3573-3580.

[88] I.S. Arvanitoyannis, M.N. Katsota, E.P. Psarra, E.H. Soufleros, S. Kallithraka, Application of Quality Control Methods for Assessing Wine Authenticity: Use of Multivariate Analysis (Chemometrics), Trends Food Sci. Techn., 10 (1999) 321-336.

[89] D.P. Makris, S. Kallithraka, A. Mamalos, Differentiation of Young Red Wines Based on Cultivar and Geographical Origin with Application of Chemometrics of Principal Polyphenolic Constituents, Talanta, 70 (2006) 1143-1152.

[90] J.W.B. Braga, C.B.G. Bottoli, I.C.S.F. Jardim, H.C. Goiciechea, A.C. Olivieri, R.J. Poppi, Determination of Pesticides and Metabolites in Wine by HPLC and Second-Order Calibration Methods, J. Chromatogr. A, 1148 (2007) 200-210.

[91] S. Perez-Magarino, M. Ortega-Heras, M.L. Gonzalez-San Jose, Z. Boger, Comparative Study of Artificial Neural Network and Multivariate Methods to the Classify Spanish DO Rose Wines, Talanta, 62 (2004) 983-990.

[92] X. Capron, J. Smeyers-Verbeke, D.L. Massart, Multivariate Determination of the Geographical Origin of Wines from Four Different Countries, Food Chem., 101 (2007) 1585-1597.

[93] F.-L. Min, Y.-W. Shi, X.-R. Liu, W.-P. Liao, HLA-B* 1502 Genotyping in Two Chinese Patients with Phenytoin-Induced Stevens-Johnson Syndrome, Epilepsy Behav., 20 (2011) 390-391.

[94] X. Hu, Z. Chen, X. Mao, S. Tang, Effects of Phenytoin and Echineacea Perpurea Extract on Proliferation and Apoptosis of Mouse Embryonic Palatal Mecenchymal Cells, J. Cell. Biochem., 112 (2011) 1311-1317.

[95] S. Khanna, K.K. Pillai, D. Vohora, Bisphosphonates in Phenytoin-Induced Bone Disorder, Bone, 48 (2011) 597-606.

[96] K. Hoshina, S. Horiyama, H. Matsunaga, J. Haginaka, Molecularly Imprinted Polymers for Simultaneous Determination of Antiepileptics in River Water Samples by Liquid Chromatography-Tandem Mass Spectrometry, J. Chromatogr. A, 1216 (2009) 4957-4962.

[97] A. Kumar, I. Xagoraraki, Human Health Risk Assessment of Pharmaceuticals in Water: An Uncertainty Analysis for Meprobamate, Carbamazepine and Phenytoin, Regul. Toxicol. Pharmacol., 57 (2010) 146-156.

[98] J.T. Yu, E.J. Bouwer, M. Coelhan, Occurance and Biodegradability Studies of Selected Pharmaceuticals and Personal Care Products in Sewage Effluent, Agric. Water Manage., 86 (2006) 72-80.

[99] S. Richardson, Water Analysis: Emerging Contaminants and Current Issues, Anal. Chem., 79 (2007) 4295-4324.

[100] K.J. Bisceglia, J.T. Yu, M. Coelhan, E.J. Bouwer, A.L. Roberts, Trace Determination of Pharmaceuticals and Other Wastewater-Derived Micropollutants by Solid Extraction and Gas Chromatography/Mass Spectrometry, J. Chromatogr. A, 1217 (2010) 558-564.

[101] S.R. Groskreutz, M.M. Swenson, L.B. Secor, D.R. Stoll, Selective Comprehensive Multidimensional Separation for Resolution Enhancement in High Performance Liquid Chromatography, Part II-Applications, J. Chromatogr. A, 1228 (2012) 41-50.

[102] S.W. Simpkins, J.W. Bedard, S.R. Groskreutz, M.M. Swenson, T.E. Liskutin, D.R. Stoll, A Versatile Tool for Quantitative Trace Analysis in Complex Matices, J. Chromatogr. A, 1217 (2010) 7648-7660.

[103] A. De Juan, Y. Vander Heydan, R. Tauler, D.L. Massart, Assessment of New Constraints Applied to the Alternating Least Squares Method, Anal. Chim. Acta, 346 (1997) 307-318.

[104] R. Bro, N.D. Sidiropoulos, Least Squares Algorithm Under Unimodality and Non-Negativity Constraints, J. Chemom., 12 (1998) 223-247.

[105] C. Tistaert, H.P. Bailey, S.C. Rutan, R.C. Allen, Y. Vander Heyden, Resolution of Spectrally Rank-Deficient Multivariate Curve Resolution:Alternating Least Squares Components in Comprehensive Two-Dimensional Liquid Chromatographic Analysis, J. Chemom., 26 (2012) 474-486.

[106] S. Peters, G. Vivó-Truyols, P.J. Marriott, P.J. Schoenmakers, Development of an Algorithm for Peak Detection in Comprehensive Two-Dimensional Chromatography, J. Chromatogr. A, 1156 (2007) 14-24.

[107] J.L. Adcock, M. Adams, B.S. Mitrevski, P.J. Marriott, Peak Modeling Approach to Accurate Assignment of First-Dimension Retention Times in Comprehensive Two-Dimensional Chromatography, Anal. Chem., 81 (2009) 6797-6804.

[108] J.C. Hoggard, W.C. Siegler, R.E. Synovec, Toward Automated Peak Resolution in Complex GC x GC-TOFMS Chromatograms by PARAFAC, J. Chemom., 23 (2009) 421-431.

[109] R. Tauler, D. Barcelo, Multivariate Curve Resolution Applied to Liquid Chromatography-Diode Array Detection, TrAC., 12 (1993) 319-327.

[110] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part A, in, Elsevier, Amsterdam, 1997, pp. 867.

[111] J.H. Zar, Biostatistical Analysis, in, Prentice Hall, Upper Saddle River, 1999, pp. 663.

[112] S.E. Reichenbach, in, 2010.

[113] A.K. Smilde, Bro, R.,Geladi, P., Multi-way Analysis with Applications in the Chemical Sciences, in, John Wiley & Sons Ltd., West Sussex, 2004.

[114] K. Horvath, J. Fairchild, G. Guiochon, Optimization Strategies for Off-Lne Two-Dimensional Liquid Chromatography J. Chromatogr. A, 1216 (2009) 2511-2518.

[115] J.V. Seeley, Theoretical Study of Incomplete Sampling of the First Dimension in Comprehensive Two-Dimensional Chromatography, J. Chromatogr. A, 962 (2002) 21-27.

[116] J.M. Davis, D.R. Stoll, P.W. Carr, Dependence of Effective Peak Capacity in Comprehensive Two-Dimensional Separations on the Distribution of Peak Capacity between the Two Dimensions, Anal. Chem., 80 (2008) 8122-8134.

[117] J.R. Whiteaker, H. Zhang, J.K. Eng, R. Fang, B.D. Piening, L.-C. Feng, T.D. Lorentzen, R.M. Schoenherr, J.F. Keane, T. Holzman, M. Fitzgibbon, Lin, H. Zhang, K. Cooke, T. Liu, D.G. Camp, L. Anderson, J. Watts, R.D. Smith, M.W. McIntosh, A.G. Paulovich, Head-to-Head Comparison of Serum Fractionation Techniques, J. Proteome Res., 6 (2006) 828-836.

[118] C.M. Shuford, A.M. Hawkridge, J.C. Burnett, D.C. Muddiman, Utilizing Spectral Counting to Quantitatively Characterize Tandem Removal of Abundant Proteins (TRAP) in Human Plasma, Anal. Chem., 82 (2010) 10179-10185.

[119] K. Peck, M.D. Morris, Optical Errors in a Liquid Chromatography Absorbance Cell, J. Chromatogr. A, 448 (1988) 193-201.

[120] D.O. Hancock, C.N. Renn, R.E. Synovec, Flow Dependence and Sensitivity of the Refractive Index Gradient Measurement with the Z-Configuration Flow Cell at Low Reynolds Number, Anal. Chem., 62 (1990) 2441-2447.

[121] W. Windig, The Use of the Durbin-Watson Criterion for Noise and Background Reduction of Complex Liquid Chromatography/Mass Spectrometry Data and a New Algorithm to Determine Sample Differences, Chemom. Intell. Lab. Sys., 77 (2005) 206-214.

[122] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagie, J. R.M., B.W. Wright, R.E. Synovec, Fisher Ratio Method Applied to Third-Order Separation Data to Identify Significant Chemical Components of Metabolite Extracts, Anal. Chem., 78 (2006) 5068-5075.

[123] N. Salkind, Encyclopedia of Research Design, Sage Publications, Inc., Thousand Oaks, CA, 2010.

[124] G.B. Limentani, M.C. Ringo, F. Ye, M.L. Bergquist, E.O. MCSorley, Beyond the t-Test: Statistical Equivalence Testing, Anal. Chem., 77 (2005) 221A -226A.

[125] R.C. Allen, S.C. Rutan, Investigation of Interpolation Techniques for the Reconstruction of the First Dimension of Comprehensive Two-Dimensional Liquid Chromatography–Diode Array Detector Data, Anal. Chim. acta, 705 (2011) 253-260.

[126] R.C. Allen, S.C. Rutan, Semi-Automated Alignment and Quantification of Peaks Using Parallel Factor Analysis for Comprehensive Two-Dimensional Liquid Chromatography–Diode Array Detector Data Sets, Anal. Chim. Acta, 723 (2012) 7-17.

APPENDIX

The three Matlab files, IKSFA_ALS_ssel.m, modcompare.m and FisherRatio.m are all written as scripts and are detailed within each code and described below. These m.files and and other associated files can be found on the R drive as follows R:\\CHEM\Rutan_lab\Hope_dissertation

IKSFA-ALS-ssel.m (m file used specifically in the analysis of urine data in Chapters 5 and 7; data sizes modified for phenytoin and wine data in Chapters 6 and 8 respectively) This script loads three chromatograms at a time and then sections them due to an out of memory issue with Matlab. Once the data is loaded, SVD (number of component determination) and IKSFA (spectral initial guess) are performed. An initial MCR-ALS (curve resolution) is done to determine appropriate constraints which are applied to the final ALS step.

The m file below calls up IKSFA and ALS4D for the determination of the number of spectral components, creation of a spectral initial guess for MCR-ALS resolution. These m.files can be found on the R drive as follows R:\\CHEM\Rutan_lab\Hope_dissertation\als4D

```
nexp=14;
s=1;
p=264;  %1st and 2nd dimension parameters for section 1
h=424;
j=11;
c=36;

%time over both chrom dimensions (30 min - 44100 points)
%time1 over "short" chrom dimension (21 sec.)
%time2 over "long" chrom dimension (30 min.)
%waves established below after loading
%using waves=dso.axisscale'{1};

time=[0:.0004166653:29.9999];
time1=[.37998:0.35:29.779484];
time2=[0:0.025:20.9999];

%load,reshape and section
%Dwight urine samples 1-3
load drs1
drs1=dso.data';
timedrs1=dso.axisscale{2};
waves=dso.axisscale{1};
clear timedrs1

load drs2
drs2=dso.data';

load drs3
drs3=dso.data';

drs1rs=reshape(drs1(912:71471,:),1,840,84,126);
drs2rs=reshape(drs2(912:71471,:),1,840,84,126);
drs3rs=reshape(drs3(912:71471,:),1,840,84,126);
```

```
X1=drs1rs(:,p:h,j:c,:);
X2=drs2rs(:,p:h,j:c,:);
X3=drs3rs(:,p:h,j:c,:);

eval(['section' num2str(s) '=[X1;X2;X3];']);

clear drs1 drs1rs drs2 drs2rs drs3  drs3rs
clear X1 X2 X3 dso

%load, reshape and section. Add this segment to section 1
%Dwight urine sample 4-6
load drs4
drs4=dso.data';

load drs5
drs5=dso.data';

load drs6
drs6=dso.data';

drs4rs=reshape(drs4(912:71471,:),1,840,84,126);
drs5rs=reshape(drs5(912:71471,:),1,840,84,126);
drs6rs=reshape(drs6(912:71471,:),1,840,84,126);

X1=drs4rs(:,p:h,j:c,:);
X2=drs5rs(:,p:h,j:c,:);
X3=drs6rs(:,p:h,j:c,:);

eval(['section' num2str(s) '=[section' num2str(s) ';X1;X2;X3];']);

clear drs4  drs4rs drs5  drs5rs drs6  drs6rs
clear X1 X2 X3 dso

%load, reshape and section. Add this segment to section1
%Dwight urine samples 7-9
load drs7
drs7=dso.data';

load drs8
drs8=dso.data';

load drs9
drs9=dso.data';

drs7rs=reshape(drs7(912:71471,:),1,840,84,126);
drs8rs=reshape(drs8(912:71471,:),1,840,84,126);
drs9rs=reshape(drs9(912:71471,:),1,840,84,126);

X1=drs7rs(:,p:h,j:c,:);
X2=drs8rs(:,p:h,j:c,:);
X3=drs9rs(:,p:h,j:c,:);

eval(['section' num2str(s) '=[section' num2str(s) ';X1;X2;X3];']);
```

```
clear drs7  drs7rs drs8  drs8rs drs9  drs9rs
clear X1 X2 X3 dso

%load, reshape and section.  Add this segment to section1
%Dwight urine samples 10-13
load drs10
drs10=dso.data';

load drs11
drs11=dso.data';

load drs12
drs12=dso.data';

drs10rs=reshape(drs10(912:71471,:),1,840,84,126);
drs11rs=reshape(drs11(912:71471,:),1,840,84,126);
drs12rs=reshape(drs12(912:71471,:),1,840,84,126);

X1=drs10rs(:,p:h,j:c,:);
X2=drs11rs(:,p:h,j:c,:);
X3=drs12rs(:,p:h,j:c,:);

eval(['section' num2str(s) '=[section' num2str(s) ';X1;X2;X3];']);

clear drs10 drs10rs drs11 drs11rs drs12 drs12rs
clear X1 X2 X3 dso

%% Load, reshape and section
% Dwight urine samples 13-14
load drs13
drs13=dso.data';

load drs14
drs14=dso.data';

drs13rs=reshape(drs13(912:71471,:),1,840,84,126);
drs14rs=reshape(drs14(912:71471,:),1,840,84,126);

X1=drs13rs(:,p:h,j:c,:);
X2=drs14rs(:,p:h,j:c,:);

eval(['section' num2str(s) '=[section' num2str(s) ';X1;X2];']);

clear drs13 drs13rs drs14 drs14rs
clear X1 X2 dso
%% PERMUTE SECTION
eval(['section' num2str(s) '=[permute(section' num2str(s) ',[2,3,1,4])];']);
eval(['chunk=[section' num2str(s) '];']);

%% LOAD CHUNK
load chunk
clear peak22 peak22rs
```

```
P=2;      % the peak of interest for area determination
e=1;     %% 1st and 2nd dimension parameters for the above selected peak
i=100;
o=19;
u=28;

eval(['peak' num2str(P) '=chunk(e:i,o:u,:,:);']);
eval(['[w,x,y,z]=size(peak' num2str(P) ');']);
eval(['peak' num2str(P) 'rs=reshape(peak' num2str(P) ',w*x*y,z);']);

figure
subplot(231)
eval(['contour(peak' num2str(P) '(:,:,1,5),70)']);  %change peak #
subplot(232)
eval(['contour(peak' num2str(P) '(:,:,3,5),70)']);
subplot(233)
eval(['contour(peak' num2str(P) '(:,:,7,5),70)']);
subplot(234)
eval(['contour(peak' num2str(P) '(:,:,9,5),70)']);
subplot(235)
eval(['contour(peak' num2str(P) '(:,:,11,5),70)']);
subplot(236)
eval(['contour(peak' num2str(P) '(:,:,14,5),70)']);

%% SVD to determine a possible starting point for iksfa analysis
clear r s* v
eval(['[r,s' num2str(P) ',v]=svd(peak' num2str(P) 'rs,0);']);
figure
plot(log10(diag(s5)),'*'); % input approriate subchunk number relative to the
soutput from the svd analysis
eval(['title(''[svd of section 1 peak' num2str(P) ']'');']);

%% IKSFA ANALYSIS

clear brow22 maxdet22 IG22
n=8;  % the number of factors to test
eval(['[brow' num2str(P) ',maxdet' num2str(P) ']=ikfsa(peak' num2str(P)
'rs,n);']);

eval(['IG' num2str(P) '=peak' num2str(P) 'rs(brow' num2str(P) ',:)'';']);


%% Initial ALS for ssel determination

pause off
figure
eval(['[first' num2str(P) ',second' num2str(P) ']=als4d(peak' num2str(P)
'rs,IG' num2str(P) ',20,.001,1);']);

eval(['ufchrom' num2str(P) '=reshape(first' num2str(P) ',w,x,y,n);']);

figure;
subplot (241);
eval(['contour(ufchrom' num2str(P) '(:,:,7,1),20)']);
title('[component 1]');
```

```
subplot (242);
eval(['contour(ufchrom' num2str(P) '(:,:,7,2),20)']);
title('[component 2]');
subplot (243);
eval(['contour(ufchrom' num2str(P) '(:,:,7,3),20)']);
title('[component 3]');
subplot (244);
eval(['contour(ufchrom' num2str(P) '(:,:,7,4),20)']);
title('[component 4]');
subplot (245);
eval(['contour(ufchrom' num2str(P) '(:,:,7,5),20)']);
title('[component 5]');
subplot (246);
eval(['contour(ufchrom' num2str(P) '(:,:,7,6),20)']);
title('[component 6]');
subplot (247);
eval(['contour(ufchrom' num2str(P) '(:,:,7,7),20)']);
title('[component 7]');
subplot (248);
eval(['contour(ufchrom' num2str(P) '(:,:,7,8),20)']);
title('[component 8]');

figure;
subplot (241);
eval(['contour(ufchrom' num2str(P) '(:,:,7,9),20)']);
title('[component 9]');
subplot (242);
eval(['contour(ufchrom' num2str(P) '(:,:,7,10),20)']);
title('[component 10]');
subplot (243);
eval(['contour(ufchrom' num2str(P) '(:,:,7,11),20)']);
title('[component 11]');
subplot (244);
eval(['contour(ufchrom' num2str(P) '(:,:,7,12),20)']);
title('[component 12]');
subplot (245);
eval(['contour(ufchrom' num2str(P) '(:,:,7,13),20)']);
title('[component 13]');
subplot (246);
eval(['contour(ufchrom' num2str(P) '(:,:,7,14),20)']);
title('[component 14]');
subplot (247);
eval(['contour(ufchrom' num2str(P) '(:,:,7,15),20)']);
title('[component 15]');
subplot (248);
eval(['contour(ufchrom' num2str(P) '(:,:,7,16),20)']);
title('[component 16]');

clear j jj
for mm=1:8   % # of components
for nn=1:16  % # of experiments
[ii,jj(nn,mm)]=max(max(squeeze(ufchrom2(:,:,nn,mm))')); %change data # by
hand to equal current t
[i,j(nn,mm)]=max(max(ufchrom2(:,:,nn,mm)));  %change data # by hand to equal
current t
end
```

```
end

%% plot of all experiments in assigned section

figure
for f=1:16
subplot (4,4,f);
eval(['contour(ufchrom' num2str(P) '(:,:,f,2),20)']); % change dimensions and
component to be plotted
end

%% Final ALS with ssel
pause off
clear ssel
ssel=NaN(126,n);
ssel(60:126,[2 3 5 6 7 8])=0; % [change to correspond to the analytes that
are not the background]

figure
eval(['[firsts' num2str(P) ',seconds' num2str(P) ']=als4d(peak' num2str(P)
'rs,IG' num2str(P) ',300,.0000001,1,[0 1 1 0 1 1 1 1],0,0,0,0,ssel);']);

eval(['ufchrom' num2str(P) '=reshape(firsts' num2str(P) ',w,x,y,n);']);

figure;
subplot (241);
eval(['contour(ufchrom' num2str(P) '(:,:,7,1),10)']);
title('[component 1]');
subplot (242);
eval(['contour(ufchrom' num2str(P) '(:,:,7,2),10)']);
title('[component 2]');
subplot (243);
eval(['contour(ufchrom' num2str(P) '(:,:,7,3),10)']);
title('[component 3]');
subplot (244);
eval(['contour(ufchrom' num2str(P) '(:,:,7,4),10)']);
title('[component 4]');
subplot (245);
eval(['contour(ufchrom' num2str(P) '(:,:,7,5),10)']);
title('[component 5]');
subplot (246);
eval(['contour(ufchrom' num2str(P) '(:,:,7,6),20)']);
title('[component 6]');
subplot (247);
eval(['contour(ufchrom' num2str(P) '(:,:,7,7),20)']);
title('[component 7]');
subplot (248);
eval(['contour(ufchrom' num2str(P) '(:,:,7,8),20)']);
title('[component 8]');

%% find area in a lope

peaks=reshape(firsts,w*x,y,n);
for m=1:14
    total(m,1)=findareabetter(peaks(:,m,2));
```

```
end

%% find 1st and 2cd max for each slice of unanalyzed data
P=9;        % the peak of interest for area determination
e=125;      %% 1st and 2nd dimension parameters for the above selected peak
i=175;
o=23;
u=27;

eval(['peak' num2str(P) '=Xfinal(e:i,o:u,:,:);']);
eval(['[w,x,y,z]=size(peak' num2str(P) ');']);

figure
for f=1:16
subplot (4,4,f);
eval(['contour(peak' num2str(P) '(:,:,f,5),20)']); % change dimensions and
component to be plotted
end

clear j jj
for mm=1:6  % # of components
for nn=1:16  % # of experiments
[ii,jj(nn,mm)]=max(max(squeeze(peak8(:,:,nn,mm))')); %change data # by hand
to equal current t
[i,j(nn,mm)]=max(max(peak8(:,:,nn,mm)));  %change data # by hand to equal
current t
end
end
```

Wine Data (the following 2 m files were used specifically for the analysis of wine described in Chapter 8)

FisherRatio.m The within class mean of the sample dimension is found and divided by the overall mean for all samples. There is an intensity weighting factor that can be applied in the last few lines of this script. It was used to for all the analysis of Chapter 8. The highest Fisher ratios are those corresponding to peaks with the most significant differences in concentration.

```
% Fisher Ratio Method as described by pierce et al. Anal Chem
% 78,14,2006,5068-5075
% data should have the following dimension order:
% injections x wavelength x chromatograms
% Xfinal is the data after the above alignment has been preformed

load Xfinal
size(Xfinal)
Xfinal=Xfinal(:,:,[1:9 11:16],:);
[w,x,y,z]=size(Xfinal);
Xfinalrs=reshape(Xfinal,w*x,y,z);
data=permute(Xfinalrs,[2 3 1]);
[h,p,b]=size(data);

n=3;    % # of measurements in ith class (group)
k=5;    % # of classes

clear ovmean meancl
ovmean=squeeze(mean(data));

meancl(:,:,1)=squeeze(mean(data(1:3,:,:)));
meancl(:,:,2)=squeeze(mean(data(4:6,:,:)));
meancl(:,:,3)=squeeze(mean(data(7:9,:,:)));
meancl(:,:,4)=squeeze(mean(data(10:12,:,:)));
meancl(:,:,5)=squeeze(mean(data(13:15,:,:)));

clear sumcl sigmacc
sumcl=zeros(126,b);

for m=1:k
    sumcl(:,:)=(sumcl(:,:)+(((meancl(:,:,m)-ovmean)).^2)*n);
end
 sigmacc=sumcl/(k-1);

 clear sumwith sigmnawith count
 sumwith=zeros(126,b);
 count=0;
 for m=1:k
     for q=1:n
     count=count+1;
     sumwith=sumwith+(squeeze(data(count,:,:))-meancl(:,:,m)).^2;
     end
 end
 sigmawith=sumwith/(h-k);
```

```
 FRwo=sigmacc./sigmawith; %without intensity weighting
 FRsumwo=squeeze(sum(FRwo));
 FRpicwo=reshape(FRsumwo,265,32);
 figure
 contour(FRpicwo,50)
 title('without intensity weighting');

 FR=sigmacc./sigmawith.*ovmean; %with intensity weighting
 FRsum=squeeze(sum(FR));
 FRpic=reshape(FRsum,265,32);
 figure
 contour(FRpic,50)
 title('with intensity weighting');

clear value index valuemin indexmin
for n=1:32
[value(n,1),index(n,1)]=max(FRpic(:,n));
end
for n=1:32
[valuemin(n,1),indexmin(n,1)]=min(FRpic(:,n));
end
```

modcompare.m This script is based on Windig's COMPARELCMS_SIM found in the PLS toolbox. It calculates similarity values between 0 and 1 with 0 being completely different and 1 being exactly the same. Further discussion is found in Chapter 8.

```
%COMPARELCMS_SIMENGINE Calculational Engine for comparelcms.
%The function calculates similarity values of variables of several
%different data sets. Plotting variables with a low similarity value
%shows the variables that are different across the samples. A typical
%example is the analysis of data sets of different batches of the same
%material with the goal to extract the minor differences between the
%samples.
%INPUTS:
%data : data cube, size n_samples, n_spectra, n_variables
%filter_width : optional, filter used for smoothing of columns in order
%     to take care of minor peak shifts, default is 1 = no filtering
%h : handle for waitbar, optional
%OUTPUTS:
%y : similarity indices of the variables, size n_variables*1.
%              Low values indicate differences.
%I/O: y=comparelcms_simengine(data,filter_width)
%I/O: comparelcms_simengine demo

%See also: COMPARELCMS_SIM_INTERACTIVE

% Copyright © Eigenvector Research, Inc. 2004-2009
% Licensee shall not re-compile, translate or convert "M-files" contained
%  in PLS_Toolbox for use with any software other than MATLAB®, without
%  written permission from Eigenvector Research, Inc.
%ww
```

```
%% Modified by HPB (1/14/2011)

% Load and reshape for wine data
load(Xfinalwo)
X=reshape(Xfinalwo,265*32,15,126);
data=permute(X,[2 3 1]);

%INITIALIZATIONS

[nslabs,nrows,ncols]=size(data);

%indexb=[from excell]; aligned to sample inj 2 using peaks 1,2,5 and the
%ave. diff. %make in matlab% subchunk is 4way data such that 2nd, 1st, inj,
spec
timeyy=[1:size(subchunk,1)];
for a=1:size(subchunk,2)
   for b=1:size(subchunk,4)
       for c=1:size(subchunk,3)
           clear Xnew2;
           Xnew2=squeeze(subchunk(:,a,c,b));
           clear Xsecond;
           Xsecond(:,1)=Xnew2((indexb(c)-
min(indexb)+1):(size(timeyy,2)+indexb(c)-max(indexb)),1)';
           Xfinal(:,a,c,b)=Xsecond(:,1);
       end
   end
end
clear Xfinalrs
load Xfinal
Xfinal=Xfinal(:,:,[1:9 11:16],:);
[w,x,y,z]=size(Xfinal);
Xfinalrs=reshape(Xfinal,w*x,y,z);
data=permute(Xfinalrs,[2 3 1]);

% Load for wine simulatated data
load sim9peakno
[w,x,v,z]=size(sim9peakno);
sim9peaknors=reshape(sim9peakno,w*x,v,z);
data=permute(sim9peaknors,[3 2 1]);
[h,p,b]=size(data);

%CALCULATE SIMILARITY INDEX
[nslabs,nrows,ncols]=size(data);
mean_spec=mean(data);
mean_spec=reshape(mean_spec,nrows,ncols);
min_spec=min(data);
min_spec=reshape(min_spec,nrows,ncols);


% array1=all(mean_spec==0);%take out all zero arrays
% array2=all(min_spec==0);
% array=((array1==1)|(array2==1));
% masses_selected(array)=[];
% mean_spec(:,array)=[];
% min_spec(:,array)=[];
```

```
% data_all(:,:,array)=[];
% max_rows(array)=[];

%CALCULATE CORELATION BETWEEN MEANSPEC AND MINSPEC

m=mean(mean_spec);
m=repmat(m,nrows,1);
s=std(mean_spec);
array=(s==0);%takes care of dividing by 0;
s(array)=1;%takes care of dividing by 0;
s=repmat(s,nrows,1);
a1=(mean_spec-m)./s;


m=mean(min_spec);
m=repmat(m,nrows,1);
s=std(min_spec);
array=(s==0);%takes care of dividing by 0;
s(array)=1;%takes care of dividing by 0;
s=repmat(s,nrows,1);
a2=(min_spec-m)./s;
y=sum(a1.*a2)/nrows;


%WEIGHS THE CORRELATION COEFFICIENTS WITH LENGTHS

a=sqrt(sum(mean_spec.^2));
array=(a==0);
a(array)=1;%prevents divide by zero error;
%y=y.*sqrt(sum(min_spec.^2))./sqrt(sum(mean_spec.^2));
y=y.*sqrt(sum(min_spec.^2))./a;
y(array)=1;

yrs=reshape(y,w,x);
figure
contour(yrs,50)

% to find SI between a given SI range within the entire data set.
idiot3=reshape(yrs,w*x,1);
[m]=find(idiot3<0.6406);
idiot3(m,:)=0;
[m]=find(idiot3>0.6831);
idiot3(m,:)=0;

%to find min and max of SI plot and their 2nd Dim locations
[value,index]=min(yrs);
```

Vita

Hope Patricia Bailey was born on August 31, 1969 in Baltimore, Maryland, and is a U.S. citizen. She graduated from Mount Carmel Christian Academy, Luray, Virginia in 1988. She received her Associate of Science degree from J. Sargeant Reynolds, Richmond, Virginia in 1999 and her Bachelor of Science in Chemistry and Forensic Science from Virginia Commonwealth University, Richmond, Virginia in 2004.

Publications

H.P. Bailey, S.C. Rutan, Chemometric Resolution and Quantification of Four-Way Data Arising from Comprehensive 2D-LC-DAD Analysis of Human Urine, Chemom. Intell. Lab. Sys., 106 (2011) 131-141.

H.P. Bailey, S.C. Rutan, Factors that Affect Quantification of Diode Array Data in Comprehensive Two-Dimensional Liquid Chromatography Using Chemometric Data Analysis, J. Chromatogr. A, 1218 (2011) 8411-8422.

H.P. Bailey, S.C. Rutan, D.R. Stoll, Chemometric Analysis of Targeted 3DLC-DAD Data for Accurate and Precise Quantification of Phenytoin in Wastewater Samples, J. of Sep. Sci., 35 (2012) 1837-1843.

C. Tistaert, H.P. Bailey, S.C. Rutan, R.C. Allen, Y. Vander Heyden, Resolution of Spectrally Rank-Deficient Multivariate Curve Resolution-Alternating Least Squares Components in Comprehensive Two-Dimensional Liquid Chromatographic Analysis, J. Chemom., 26 (2012) 474-486.

H.P. Bailey, S.C. Rutan, Comparison of Chemometric Methods for the Screening of Comprehensive Two-Dimensional Liquid Chromatographic Analysis of Wine, Anal. Chim. Acta, submitted July (2012).

Presentations

Sarah E. G. Porter, Sarah C. Rutan, Hope P. Bailey, Dwight R. Stoll, Peter W. Carr, Jerry D. Cohen "Multi-way Analysis of 2D Liquid Chromatographic Metabolomics Data" poster presentation at Systems Biology Conference. Richmond, Virginia

Hope P. Bailey, Sarah C. Rutan, Dwight R. Stoll, Peter W. Carr "Multi-way Analysis of 2D Liquid Chromatographic Metabolomics Data" poster presentation at High performance Liquid Chromatography in Baltimore, Maryland, 2008.

Hope P. Bailey, Sarah C. Rutan, Dwight R. Stoll, Peter W. Carr "Quantitative Analysis of Comprehensive 2D Liquid Chromatographic (2D-LC) Data using Multivariate Techniques" poster presentation at ACS in Washington DC, 2009.

Hope P. Bailey, Sarah C. Rutan. "Issues Affecting the Analysis of Data Arising from Comprehensive 2D-LC-DAD" Oral presentation at PittCon in Orlando, Florida, 2010

Hope P. Bailey, Sarah C. Rutan. "Chemometric Analysis of Beverages Following Separation by Comprehensive Two Dimensional Liquid Chromatography with Diode Array Detection" oral presentation at PittCon in Atlanta, Georgia 2011

Hope P. Bailey, Sarah C. Rutan. "Comparison of Screening Methods for the Analysis of Four-Way LC × LC Data to Determine Concentration Differences among Wine Samples" oral presentation at CAC in Budapest, Hungary 2012